Historical Software Issue 8: The Kurzweil Data Entry Machine (KDEM)

Thaller, Manfred

software

HISTORICAL SOFTWARE SECTION[+]

Issue 8: The Kurzweil Data Entry Machine (KDEM).

As in the last of these sections, we will deal in this one with software but with an important new development in the field of hardware, the Kurzweil Data Entry Machine, a device being allegedly able to read any printed or typed matter. As will become clear from our discussion, this device is (in price and ability) definitely beyond the scope of an independent researcher in Social Historical Research. We think, that it's potential value is large enough to explain about in detail.

1. The basic principles of a KDEM (1)

A Kurzweil Data Entry Machine consists of four basic units:

- a scanner turning printed text into a preliminary digital representation,
- a mini computer converting that preliminary representation into the standard EBCDIC or ASCII codes,
- a graphics terminal giving the user control over the way all that is done,
- an intermediate random access storage unit for holding various information about the fonts being processed and the preliminary output produced from the material currently being read and
- an output facility for transmitting the final product to some other computer (either via a direct link or via a standard magnetic tape or floppy disk).

Leaving more specialized components aside, the central mini computer is provided with two major pieces of software:

- the "main" system for the interactive conversion of printed texts and
- a postprocessor allowing to manipulate further the data gained from the first program before writing it to tape (or transmitting it to another computer). This postprocessor provides specialized string handling capabilities which markedly enhance the power of the system as a whole.

As the system is intended to read all (Latin) fonts, it is obvious that it treats none of them with any particular favor (it does e. g. not do necessarily a particular good job with OCR letters). It has, on the contrary, to "learn" how to understand any given font. Reading a given specimen of print proceeds in three steps:

[+]Address all communications to: Manfred Thaller, Max-Planck-Institut für Geschichte, Hermann-Föge-Weg 11, D-3400 Göttingen.

a. The system is first calibrated. That is, it is presented with a few lines of the print desired to be understood and given such additional information as is necessary to understand how (un)evenly the types are spread over a line.

b. Next the system is trained. That is, it is given a number of additional lines to process. It will first try to read every letter according to the standard assumptions of the central software components and ask the user to confirm this reading or change it. The confirmations and/or changes to the original readings are integrated into a 'character definition table' that describes the font just being processed. So during training the number of necessary interventions of the operator should (and as a rule will) steadily decline. The character definition table can be stored at any point of training, which can be resumed after any interval of time.

c. Finally the production phase starts. This means, that the KDEM reads the material presented to it line after line, tries to identify each character encountered and computes internally a probability value for the identification. If this value lies beyond a treshold (that can be influenced by the user) the character will be accepted; if not, the operator will be asked to confirm the reading or put another character instead of it.

2. Operating speed and practical power of the system

The following comments are primarily based upon a test by the author, undertaken during May 1982 at the Compact Computer Systeme GmbH. in Hamburg (2), which is the main supplier of KDEM's for the German speaking countries. In about 8 hours working time the system was tested with about a dozen of different fonts as appearing in typical editions of historical source material. Those experiences were augmented by the evaluation of the conversion of source material done by a commercial firm mentioned below. Besides this first hand information, the author had the possibility to talk to a number of people who had tried out (or conceptualized) the KDEM for several applications. For various reasons this information is considered confidential and not ascribed individually. One should hint, though, at Susan M. Hockey of the Oxford University Computing Service, where a KDEM is used for the academic community for some time already.

2.1 Range of applicability

While the Oxford University Computing Service claims to have used a KDEM for Oriental alphabets and at the Max-Planck-Institut für Rechtsgeschichte in Frankfurt am Main one was recently acquired with the explicit aim of converting large corpora of Greek sources into machine readable form, the tests upon which this report is based primarily were undertaken with Latin alphabets. The unreflected experience is somewhat conflicting.
On the one hand one had to be astonished that the KDEM could learn to read a high proportion of a large number of fonts with astonishing speed: so after training on only 6 lines of the protocolls of the Deutsche Nationalversammlung of 1848/49, being printed in the German alphabet, 70 % of the characters were already identified correctly after about a minute. One should add, that even the distributor for

Germany was doubtful, if the system would be able to read such stuff without prolonged effort. When we used texts printed more cleanly in the very first "job" done by this author on his own, that is, without any help of experienced users, it took an hour to train the system so well, that a specimen from the Freiherr-vom-Stein Gedächtnisausgabe could be read 100% correctly.
On the other hand: at least for an unexperienced user, there remain quite a few obstacles. The KDEM is able to resolve ligatures and to recognize (in principle) characters, which consist of two separate parts. In some of the specimens used, the author was unable, however, satisfactorily to process letters "n" and "u", when pressure during print was low, so that the connecting line between the two vertical ones was not fully printed. (The lacunae between the two vertical lines can often not be seen without some magnifying lens.) Credible reports have been received,though,that this is a problem of training and such difficulties can be overcome with additional experience.
This additional experience is particularly useful when typographical refinements shall be depicted accurately. The KDEM allows to keep a very large number of characteristics of printed matter as flags within the text made machine readable, as e.g., italics, bold prints, sub- and/or superscripts within a text, tabulators and many more. Most of this features will not be necessary for applications arising from Social Historical Research, as in our field one will generally restrict oneself to the plain wording of a text and be able to ignore diacritics or style of printing - though,reading, e.g., a large amount of printed statistics, where figures actually computed are printed in the plain font, and figures implied in italics, one may come surprisingly soon to applications where such things matter.
When the tests of the author in Hamburg started, he submitted a text, containing quite a few such features. The KDEM was calibrated and trained on that text (the author being trained at the same time how to use it later in tests on his own) by the personell of the supplier. As it turned out, a feature like a change from the usual print to italics within the same word could not satisfactorily be resolved during this initial training, though it should not overcome if handled by a really experienced operator.

2.2 Operating speed

The remarks just made should make it completely clear, that everything that can be said about operating speed has to be heavily qualified. The manufacturer claims, that the training phase for a completely new font will take between 10 and 30 minutes, "possibly longer with more complicated texts". This is true if and only if the system is operated by a person that uses it continously and knows a lot about the tricks of this particular trade, which are not documented. Completely ridiculous would be the concept, that somebody without any previous experience with the system coud use it just like a XEROX machine. So, one can use a KDEM efficiently in two cases: either if large corpora of one and the same font are to be processed, or if the machine is used in an environment, where a really experienced operator is around. (People who have actually used the system for production work say that one needs at least 2

weeks really to know how to handle it - verbal communications from Oxford University Computing Service say, that it took them about a year really to run the machine at its full power. (But then one has to say, that OUCS certainly has up to now the most sophisticated applications.)

2.3 Some reflections on the experiences had

The KDEM is certainly an extremely powerful machine. It can, in principle, read all printed material (and a large number of typoscripts) historians will turn up with. It will do so only operated by an expert though. This means:

- for acquisition: there's nothing like a "benchmark". If you take the most simple text up to Hamburg (or whereever your local supplier resides) and intend to decide if the system can solve your problems, you may learn a bit about the physical and psychological state the operator there is in at the day and time of your visit, but the outcome of such a "test" says nothing about the applicability of the KDEM to the material in question. If you really want to find out, what the machine is able to do, you will either have to send larger test corpora to a commercial house (and find out how difficult a job they consider reading such material by the price they charge) or, even better, make your tests long enough to get a good working knowledge of the machine yourself; what will certainly not take less than a week.

- for production work: this is no machine a computing centre can buy, plug it in, put the manual unto a shelf and hope that the user community will somehow understand what to do. A KDEM is sensible only as either a generalized investment into the infrastructure for one or more disciplines (being situated at a given computing centre and handled by an operator there, ready to convert any material sent from any institution supported by the funding agency in question) or as part of a long range attempt at the systematic conversion of a large corpus of source material relevant for some discipline (converting during normal operation that corpus, but being available for any work somebody from the discipline requests on short notice).

3. An evaluation of the usefullness of a KDEM

We think not to have been overly enthusiastic about this technical development that far and to have shown clearly the limitations it has and the difficulties its practical use poses. So we feel to be entitled to state equally clear that we think, the potential of this device for all computerwork in particular can scarcely be overestimated. We will just quote 4 applications and leave the rest to the statement, that the real importance of the KDEM lies in the fact, that, provided a skilled operator is available, any text currently existing in print or typescript of less than, say, 500 pages can be machine readable and error checked with at least the same accuracy than with other means of input within a week or considerably less.
Possibility 1: Leaving all methodological considerations aside, it is incomparably easier to organise all applications which see information retrieval as their first and statistical analysis only as their second goal. A complete KWIC to one of the volumes of the Freiherr-vom Stein Gedächtnisausgabe (3) may even now come as cheap as 600 DM and be available in less than a week, when one asks a commercial house to convert the book into machine readable form and

has appropriate software available (4).
Possibility 2: Every content analytical approach can cut the time of data input from one or two years to a month or less, if newspaper clippings, specimen chapters from books or a set of letters (which have to be available in typescript) are not retyped via video screen (nothing to say about punched cards). Please note, that we are not speaking about sending such material to commercial houses: such collections will necessitate such frequent training phases (as the font is changing quite often) that they could be processed only if somebody from the research project to be supported gets access to the KDEM itself and is introduced into its use by an experienced operator.
Possibility 3: Every quantitative research project using printed sources which remotely resemble lists can cut the stage of data input to days and at the same time substantially increase the reliability of coding and the general quality of the data, provided suitable software for data preparation is available.
Possibility 4: When the time needed for the input of large corpora of texts shrinks drastically, it finally should become possible to bridge the gap between people using EDP for hermeneutic purposes ("only") and such which ("just") produce statistics. In my opinion still the most important development to hope for among computer users in the Humanities.

4. How far are such services available right now?

Currently there are two possibilities (from a German point of view) to access KDEM's if you pay for the conversion job done.
Specialized on applications in Literary and Linguistic Computing is the Oxford University Computing Service (5), which in principle is ready to convert any printed matter into a magnetic tape for a relatively modest payment. While very advisable for our readers in the UK, our German members will think it somewhat less than perfect to have to send their material abroad. The very recommendable thing about the OUCS: it certainly is the most experienced institution in the world, when it comes to the handling of literary texts with a KDEM and, given its history as an institution, one can with equal certainty assume that it would do everything to do a good job with any other class of material sent to it. A drawback: OUCS operates fully in concordance with the very letter of the British copyright laws. In one case OUCS wrote an applicant, who wanted to convert the poems of a less well known poet from the beginning of this century to magnetic tape, that it would in principle be willing and able to undertake the job, but needed a written permission of the copyright holder. This being a less well known poet indeed (the last facts about him that could be ascertained originating from the 1920s and the publishing house gone since even longer) the potential KDEM user despaired and entered the poems via video screen.
A good alternative (for German users probably the first choice) is the Prisma Datenerfassung in Hamburg (6). Being a commercial firm, Prisma is specialized upon the conversion of large amounts of data, where the training has to be undertaken only once. If your material falls into this category you will be served very well, at prices somewhere between DM 1.90 and DM 3.20 per thousand characters.

Prismas commercial orientation has two drawbacks: first of all, it will be reluctant to accept any material where during production work heavy editing has to be done. So attempts to read by them parts of the protocolls of the Deutsche Nationalversammlung in Frankfurt 1848/49 came to nothing. Secondly the operators there are obviously used to modern German. In principle this does not bother the KDEM, but, if you have to process which are not modern German (say Latin (tested) or (supposedly) French or a dialect) an operator who gets a letter for verification out of a word he/she does not recognize will not always make the correct decision.
To sum it up: both institutions do a good job at converting texts; none of them will be a good selection if you have to work with newspaper clippings or other selections of material, which might be converted pretty fast, if a student can sit down with them at the keyboard of a KDEM him/herself - and is advised by an experienced operator during training.

5. A final remark

After doing some work on research designs for large multi-user historical data bases himself and talking to many people about the potential value of a KDEM for their respective research projects, this reviewer thinks that many of them could profit substantially, in some cases spectacularly indeed, if such a device would be available in a suitable computing centre and accessible for, say, every research project at a German university or non-profit organisation. One of the main problems is, that it is currently extremely difficult to estimate how many users such an infrastructural investment might get. QUANTUM would appreciate therefore, to get informal communications from those members which think they would profit by it, containing: a) institutional affiliation, b) a short description of the material in question, c) a guess how much material there is to be converted and d) a short notice, if you are talking about research already in progress or about what you would do, if a KDEM would be accessible to you.

Software section has during the last issues been rather far from any statistical considerations: during the next four to six issues we will reverse this course and report about the facilities of the new versions of the "big" statistical packages (one per issue).

FOOTNOTES

1 A hardware review with somewhat more technical details (from the point of view of Literary and Linguistic Computing) may be found in Computers and the Humanities 15 (1981), 183-185.

2 Dipl.-Ing. Rolf Saacke, CCS GmbH., Schwanenwik 32, D-2000 Hamburg 76. KDEM's are also available from Dipl.-Ing. D. Sommer, Elektronik und Datenverarbeitung, Jahnstr. 49, D-6000 Frankfurt am Main.

3 Tested with Ekkehardi IV Casus St. Galli, Ed. by Hans F. Haefele, Darmstadt 1980, (= Ausgewählte Quellen zur Deutschen Geschichte des Mittelalters).

4 On rough and ready KWIC's as a substitution for "yet another retrieval system" see the Historical Software Section in HSR 23 (July 1982).

5 Susan M. Hockey, Oxford University Computing Service, 13 Banbury Road, Oxford OX2 6NN, Great Britain.

6 Günter Paff, Prisma Datenerfassung GmbH., Pflugacker 36 d, D-2000 Hamburg 54.