NBS TECHNICAL NOTE

462

Nonnumeric Data Processing in Europe

A Field Trip Report
August-October 1966

U.S. DEPARTMENT OF COMMERCE
National Bureau of Standards

NATIONAL BUREAU OF STANDARDS

The National Bureau of Standards[1] was established by an act of Congress March 3, 1901. Today, in addition to serving as the Nation's central measurement laboratory, the Bureau is a principal focal point in the Federal Government for assuring maximum application of the physical and engineering sciences to the advancement of technology in industry and commerce. To this end the Bureau conducts research and provides central national services in three broad program areas and provides central national services in a fourth. These are: (1) basic measurements and standards, (2) materials measurements and standards, (3) technological measurements and standards, and (4) transfer of technology.

The Bureau comprises the Institute for Basic Standards, the Institute for Materials Research, the Institute for Applied Technology, and the Center for Radiation Research.

THE INSTITUTE FOR BASIC STANDARDS provides the central basis within the United States of a complete and consistent system of physical measurement, coordinates that system with the measurement systems of other nations, and furnishes essential services leading to accurate and uniform physical measurements throughout the Nation's scientific community, industry, and commerce. The Institute consists of an Office of Standard Reference Data and a group of divisions organized by the following areas of science and engineering:

Applied Mathematics—Electricity—Metrology—Mechanics—Heat—Atomic Physics—Cryogenics[2]—Radio Physics[2]—Radio Engineering[2]—Astrophysics[2]—Time and Frequency.[2]

THE INSTITUTE FOR MATERIALS RESEARCH conducts materials research leading to methods, standards of measurement, and data needed by industry, commerce, educational institutions, and government. The Institute also provides advisory and research services to other government agencies. The Institute consists of an Office of Standard Reference Materials and a group of divisions organized by the following areas of materials research:

Analytical Chemistry—Polymers—Metallurgy — Inorganic Materials — Physical Chemistry.

THE INSTITUTE FOR APPLIED TECHNOLOGY provides for the creation of appropriate opportunities for the use and application of technology within the Federal Government and within the civilian sector of American industry. The primary functions of the Institute may be broadly classified as programs relating to technological measurements and standards and techniques for the transfer of technology. The Institute consists of a Clearinghouse for Scientific and Technical Information,[3] a Center for Computer Sciences and Technology, and a group of technical divisions and offices organized by the following fields of technology:

Building Research—Electronic Instrumentation — Technical Analysis — Product Evaluation—Invention and Innovation — Weights and Measures — Engineering Standards—Vehicle Systems Research.

THE CENTER FOR RADIATION RESEARCH engages in research, measurement, and application of radiation to the solution of Bureau mission problems and the problems of other agencies and institutions. The Center for Radiation Research consists of the following divisions:

Reactor Radiation—Linac Radiation—Applied Radiation—Nuclear Radiation.

---

[1] Headquarters and Laboratories at Gaithersburg, Maryland, unless otherwise noted; mailing address Washington, D. C. 20234.
[2] Located at Boulder, Colorado 80302.
[3] Located at 5285 Port Royal Road, Springfield, Virginia 22151.

UNITED STATES DEPARTMENT OF COMMERCE
C. R. Smith, Secretary
NATIONAL BUREAU OF STANDARDS • A. V. Astin, Director

TECHNICAL NOTE 462

ISSUED NOVEMBER 1968

Nonnumeric Data Processing in Europe

A Field Trip Report
August–October 1966

Mary Elizabeth Stevens

Center for Computer Sciences and Technology
Institute for Applied Technology
National Bureau of Standards
Washington, D.C. 20234

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C., 20402 - Price 65 cents

CONTENTS

FIGURES

NONNUMERIC DATA PROCESSING IN EUROPE: A FIELD TRIP REPORT

MARY ELIZABETH STEVENS

A number of nonnumeric data processing projects in the United Kingdom, Belgium, the Netherlands, Italy, Sweden, the Federal Republic of Germany, and the U.S.S.R. have been visited. Topics covered include character and pattern recognition; speech analysis, synthesis, and recognition; artificial intelligence; mechanized documentation and library automation; linguistic data processing, and computing and programming theory.

Key Words: artificial intelligence, automatic abstracting, automatic indexing, computational linguistics, computer centers, documentation, information storage, selection and retrieval, library automation, optical character recognition, pattern recognition, programming languages, speech and speaker recognition.

1. INTRODUCTION

With the support of the Information Systems Branch, Office of Naval Research, a field trip survey of selected nonnumeric data processing research and development projects in the United Kingdom, Belgium, Sweden, the Netherlands, Italy, and the Federal Republic of Germany was conducted during August and September of 1966. In addition, under the auspices of UNESCO, the author participated in a Symposium on Mechanized Abstracting and Indexing held in Moscow, U.S.S.R., September 28 - October 1, 1966.

The field of "nonnumeric data processing" is both broad and varied, including, for example, computational linguistics and linguistic data processing; character reading and pattern recognition; library automation and mechanized information selection, storage and retrieval systems; speech synthesis, analysis and recognition; and research in cybernetics, self-organizing systems and artificial intelligence. In this report, the area of optical character recognition will be considered first, followed by that of speech analysis, synthesis, and recognition, and of automatic pattern recognition more generally. Information storage, selection and retrieval, library automation, and mechanized documentation projects will then be considered. Next, some of the European work in computational linguistics and linguistic data processing will be reviewed. Computer centers, computing and programming theory, artificial intelligence, and miscellaneous other projects will constitute a final discussion area.

It should be noted that certain commercially available equipment instruments, materials, and computer programs are identified in this report in order to describe adequately techniques used in some of the various projects that were visited. In no case does such identification imply either favorable or unfavorable evaluation nor endorsement or criticism by the National Bureau of Standards.

2. OPTICAL CHARACTER RECOGNITION

European projects in optical character recognition that were visited included those of International Computers and Tabulators, Ltd., the General Post Office, and the National Physical Laboratory in the United Kingdom; those of Telefunken, Standard Elektrik Lorenz, Siemens and Halske, and the Technische Hochschule, Karlsruhe, in the Federal Republic of Germany; and those of Olivetti, Milan, Italy, and the Laboratory for Electromodelling, Academy of Sciences of the U.S.S.R., Moscow.

Considerations of relatively limited markets as well as limited resources tend to explain typical West European emphasis in the development and application of optical character recognition techniques on small, relatively low speed, inexpensive devices and systems. A related emphasis is upon high reliability, especially in connection with standardized and stylized fonts, but also directed to the development of capabilities for handling a wide variety of relatively poor quality materials.

There appears to be comparatively greater emphasis in European efforts as contrasted to those efforts in the United States to insist upon a more conventional and aesthetically pleasing character set for standardized fonts --- that is, the OCR-B recommended font rather than OCR-A. It is to be noted in particular that West European manufacturers, through the European Computer Manufacturers Association, contributed financial support to the development of prototype character styles that have resulted in the more conventional legibility of OCR-B.

The most directly operational applications of character recognition techniques in Western Europe are those of the automatic reading of well-controlled, highly stylized, numeric and alphanumeric character sets. Such alphanumeric information is typically read by machine for accounting, banking, mail sorting, and other management or business information processing purposes. In particular, the automatic reading of turnaround documents imprinted with standardized font characters provides a practical area for use of both magnetic ink (MICR) and optical character (OCR) recognition equipment.

2.1 United Kingdom

A first European example of limited, stylized font character reading is provided by the continuing operation of two Solartron installations in the United Kingdom. Although Solartron is no longer engaged in manufacturing character recognition equipment, one of its installations is a mark reader utilizing CRT scanning of football pool betting forms in order to compute the values of coupons and to sort them. In a thirty-hour period each weekend in the season, up to a million documents are automatically processed. The second Solartron system that is still operating is at Montague-Burtons in Leeds. This is an optical tally-roll reader for a limited numeric font only, and it converts from tally register paper tape rolls to punched cards. (Fig. 1).

At the Stevenage establishment of International Computers and Tabulators, Ltd., Dr. Peter Hall is manager of the character recognition development program. Although ICT patents in the field apparently date back prior to World War II, the current activities started only about four years ago. However, the staff represents considerable prior experience at Solartron, Farrington, and Rank-Xerox.

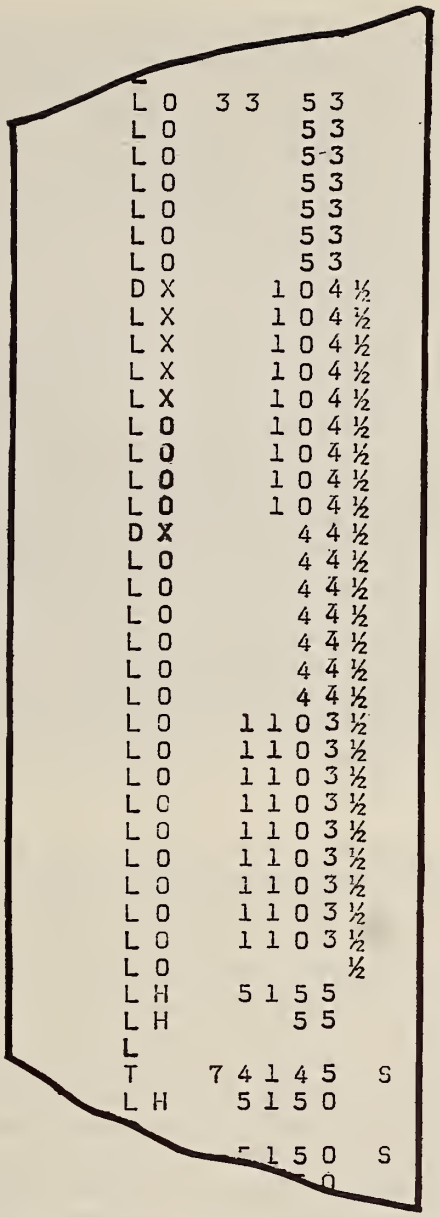


Figure 1. Sample of Tally Roll Record Read by Machine

ICT began its recent activities by entering into agreements with other manufacturers, e.g., with the National Data Processing Corporation for purchase of MICR check readers and also of a Readatron optical machine. The latter was used only for test purposes, there having been a poor market for credit card applications in the United Kingdom at that time. About a dozen National Cash Register check-sorters have been purchased for re-sale and have apparently been quite successful, particularly at the Swedish Bank Girocentralon. Another NCR machine is being modified for the ICT Standard Interface equipment and will subsequently be installed at the Bank of England.

In terms of long-range interests in optical character recognition, ICT policy has stressed the need for standard fonts and for document transport developments specifically designed for character recognition applications. The company participates actively in the work of the European Computer Manufacturers Association, and the recognition development work at ICT has centered largely on the International Standards Organization's OCR-B alphanumeric font. However, ICT has first designed a mark reader with 24 photocell tracks (any one of which may be selected as clock, the other 23 being used for data recorded with ordinary pencil) and with computer-controlled format flexibility.

Design specifications for the OCR-B font line reader were established in the early 1960's, preproduction models have been built, and the machine was to have been formally demonstrated in October 1966, with next-stage production models following. The specifications for transport and handling emphasize flexibility for a wide range of potential markets. Varying document sizes from 8 1/2" x 13" down to 3" x 5" card size will be accommodated at 300 items per minute (high speed is not deemed particularly important for the market that ICT is aiming toward), with capability for handling longer documents at reduced speeds. The handling equipment has been specifically designed so that different kinds of reader-recognition devices can be connected to it. For example, they anticipate applications in which the OCR-B line reader can be used simultaneously with an optical mark reader.

ICT chose a drum reading system because of the drum's advantages of (1) providing a stable reading platform and close control of speed, (2) having re-read facilities, (3) maintaining continuous scope display of re-read documents, and (4) serving as intermediate storage for cases of delay in computer access, especially for multiprogrammed or time-shared machine systems.

The OCR-B machine is designed as a single-line reader, with facilities for repositioning to read a line anywhere on the document. (Some work has been done on page reading, e.g., feeding from the document transport sidewise onto the drum and reading one line per drum revolution, but there is no ICT interest in multifont techniques at this time because of their commitment to the standard font). The page reader will require further work with a CRT scanner, still under development.

For the line reader, a Vidicon scanning head, similar to the RCA Videoscan but with what is claimed to be improved resolution, is being used. The recognition logic is similar to that of the Readatron with the input pattern being shifted through a register and systematically sampled by a 14 x 22 resistor correlation matrix. Conservative processing of the character involves about 75 percent sampling of the input image. Computer techniques were used to set up appropriate values for the correlation matrix. Digital methods are used for stroke width normalization, with preprocessing of one, two, or three cell widths through appropriate logical networks in order to output a three-cell stroke. Performance indications for numeric characters to date are claimed to be very good, and certainly the results are better than typical MICR performance.

An important ICT development is the ICT Print Quality Monitor. (Fig. 2) [1]. This equipment involves a CRT scanner and light feedback loop; it displays variable reflectance values, and it makes measurements of character quality automatically. Another ICT project involves the development of a 60-photocell scanner for high-speed check sorting (i.e., paper speed of 400 inches per second) but some problems with the recognition logic have been encountered at this speed.

The problems of mail sorting and handling as partially or fully mechanized operations are a continuing concern of the British General Post Office. At the Dollis Hill Research Station of the Engineering Department, General Post Office, a team headed by Dr. A. W. M. Coombs has been investigating the variety of practices and styles with respect to the

Figure 2. I.C.T. Optical Print Quality Monitor
(Photograph courtesy of International Computers Limited)

addressing of mail and the typical noise to be encountered in representative samples of mail.

In London, a postal coding system has been used for many years. The scheme is somewhat more elaborate than the American Zip Code. There are approximately 140 postal zones in the London metropolitan area, identified by alphanumeric symbol sequences that are deeply engrained in public use, such as "N. W. 2" or "E. C. 4", with or without punctuation. Experiments at Norwich have involved the initial steps toward the formulation of a nation-wide postal address code. In the proposed general system, the postal code sequence will be comprised of two groups of laphanumeric code sequences, the first group indicating the post "town" and the second indicating a subdivision that corresponds to a designated portion of a postman's "walk".

The mail sorting operations are carried out manually at the present time. These operations involve the reading and keystroking of both sections of the postal code, and phosphorescent binary markings are produced in order to control between three and six subsequent machine sorts per piece of mail. The immediate objective of the GPO character recognition research is therefore to replicate the manual recognition and marking operations with the following goals: (1) one in 2, 000 item accuracy tolerance (there being no redundancy in the codes), (2) no more than three errors per 1, 000 for the outward part of the code, (3) investment costs for the reader equipment of the equivalent of $50, 000, and (4) processing rates at 500 pieces of mail per minute.

The problems to be solved include, first, the variety of types of address inscriptions that are typically encountered. Only 56 percent or less of letter addresses are typed, there is a heavy proportion of handwritten address inscriptions, and even when the addresses are typed the quality is generally poor. A second problem area is that of the identification of the code itself wherever it may be embedded in the full address. "It rapidly, becomes obvious, when 'Live' mail is studied, that recognition of the letters and figures constituting

the code is only a part of the problem. The code itself is often so embedded within the address, or so surrounded by other matter, that it is very difficult to formulate a set of rules which a machine could follow even to find it." [2]

Fact-finding studies conducted to date by the GPO have included the photographing of samples of live mail. The face images of envelopes of approximately 20,000 pieces have been recorded on 35mm film, which is subsequently scanned and converted to 5- or 8-channel paper tape. The quality of the original address appearance can be re-displayed and is quite poor, being far below the limits of available character-reading equipment capable of operating at an acceptable reject rate.

Since stylized or standardized fonts are generally unacceptable solutions to the mail sorting problem, the basic GPO approach is therefore that of adaptive and self-adjusting majority-logic pattern recognition, following the concepts of Rosenblatt's multi-layered "Perceptrons" in part, and also those of "Conflex I" of Scope, Inc. In the case of 20 multifont examples, each, of mutilated members of a 10-character vocabulary, the GPO researchers were able to develop a technique of "shuggling" (or shuffling or jittering) of the input image for maximum and minimum correlata with the same and with other character patterns in the vocabulary. The results were subsequently screened by human eye, to provide a reduced and somewhat irregular (roughly, 20 x 10) matrix of criterial points for character discrimination. They subsequently recognized successfully 198 of 200 source pattern inputs. In general, the GPO plans to develop its own prototype reading recognition equipment. It is estimated that a prototype in the 500 document/minute speed range can be ready in three years. Some advance simulations have been carried out on their computer.

At the National Physical Laboratory, in Teddington, there are a number of projects in character and pattern recognition research, as well as other nonnumeric data processing research and development activities. The character and pattern recognition work ranges from developments aimed toward short range practical applications to longer range experimental investigations. In particular, the CYCLOPS machine, based on the J.R. Parks recognition system (an electronic version of detecting the criterial areas of character auto-correlations as developed optically by M. Clowes [3], [4]) reads printed numerics, at 3,000 characters per second, and 10 characters per line, for tally roll reading applications. The Plessey Company has taken over this design in order to develop it for commercial production.

In the CYCLOPS design, analog techniques are used to carry gray-scale information along to the recognition logic. Position-invariant n-tuple operations on quite low quality printed numerics are used to detect typical features. Because of integration, various translations of the numeric character in the field of view can be accommodated, so that there is considerable tolerance for skew. Specifications for the Plessey commercial machine include the reading of OCR-B size 1 numerics at 1,200 characters per second.

M.O. Clayden and J.R. Parks of the National Physical Laboratory are also continuing research on more difficult problems of character and pattern recognition [5]. A multi-layer hierarchical system is currently under investigation that applies some of the original ideas of auto-correlation but that no longer integrates. Instead, both positive and negative copies of the input image are superimposed and are shifted in various directions and with special screens to provide mappings that discriminate the criterial features. In the second level of operations, coincidences are detected for relative dispositions of these features in the characters making up the alphabet. For example, "h", "u", and "n" will have the same parallel-verticals feature, the possibility of "u" will be eliminated by a mapping for mid-area curves, and a long-ascender feature will show "h" rather than "n", or, conversely, the negative of the long-ascender feature will detect "n".

It is to be noted in particular that the input image for these NPL experiments may be a sequence of characters, such as a complete word of text, with or without spacing between its individual characters. Thus, prior segmentation is not required because the technique, in effect, scans the text through a moving aperture and checks continuously for specific groupings of local features that will identify the characters one by one. In this connection, simulation experiments have been run on the Ace computer for all possible pairs of 10 alphabetic characters with no space between them. It has been possible to detect 12 different local features in the first level of operations and 4 different possible definitions for each character at the second level of operations. While false recognitions occur (e.g., two quantized "ow" 's being detected at one stage as an "x"), the results appear promising in general. Further planned research will apply threshold requirements with respect to clumps of the most probable character values. Statistical data from tests to date is available as to the extent by which recognition improves as spacing or segmentation is reintroduced.

2.2 Federal Republic of Germany

In Western Germany, about seven years ago, Telefunken AG entered into a contract with the German Post Office for experiments and pilot mechanization of the German postal check processing operations, especially the receipt, accounting, statement updating, and credit-debit transfer operations. In general, paper-handling requirements for this application are far more rigorous than any of the comparable magnetic ink or optical character recognition applications in the United States. Envelope opening is required at input, and output involves envelope stuffing with both a variable number of transaction documents and a computer print-out of the revised statement of the account balance. [6]

Starting in 1958, an early decision was reached by the German Post Office to proceed with MICR --- first, the American E13B font, then the Bull CMC-7 --- and the first two operating installations of the Telefunken model BSM 1050 sorter (Fig. 3) were for magnetic ink applications. However, this equipment is now being adapted for reading of OCR-A font numerics.

The booking unit of the BSM 1050 Document Sorting Machine consists of a reader, 16 output pockets, and 2 input stackers. The machine is computer-controlled to read on demand from either of the two input stations. The input from both of these stations may also be merged as required. The present system requires magnetic tape storage of the basic account information. Thus, waiting stations may be required. The reader-mechanism therefore picks up one document, reads it onto a drum, and holds it until the account has been found. Then only is the next document from this input station called up.

The first sorting of input items is to account number, since there are as yet no random access facilities for account lookup in the system. The provision of a second input station also allows the special handling of urgent items, such as withdrawals made a few minutes before closing time. On-line bookkeeping results in a revised account statement that is prepared on an Analex printer. The pertinent documents are merged and are packed together with the new statement into window envelopes. Future developments will involve consideration of possibilities for printing onto the envelope directly and for the automatic reading of the German postal district codes.

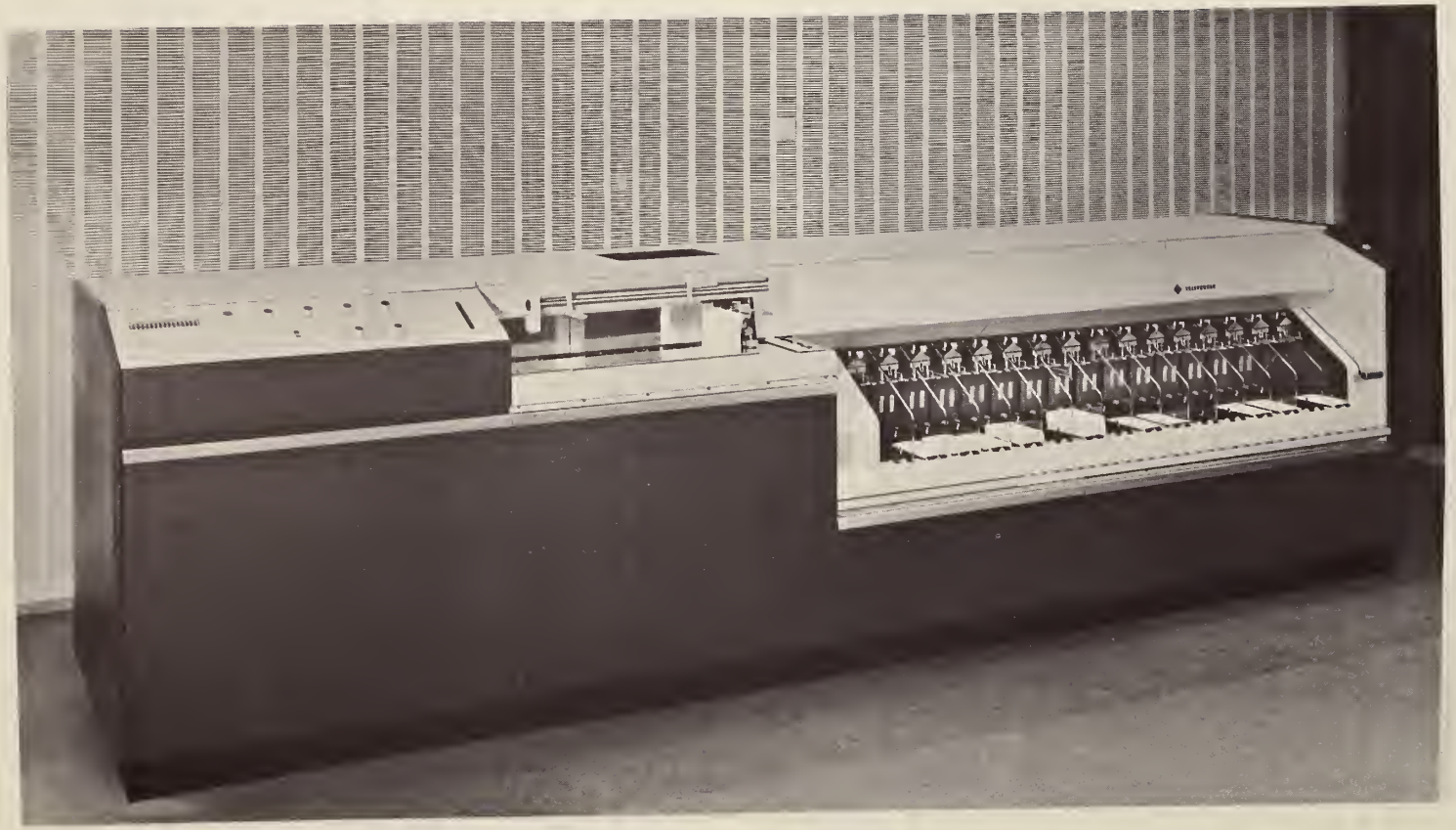


Figure 3. Telefunken BSM 1050 Document Sorting Machine
(Photograph courtesy of AEG-Telefunken)

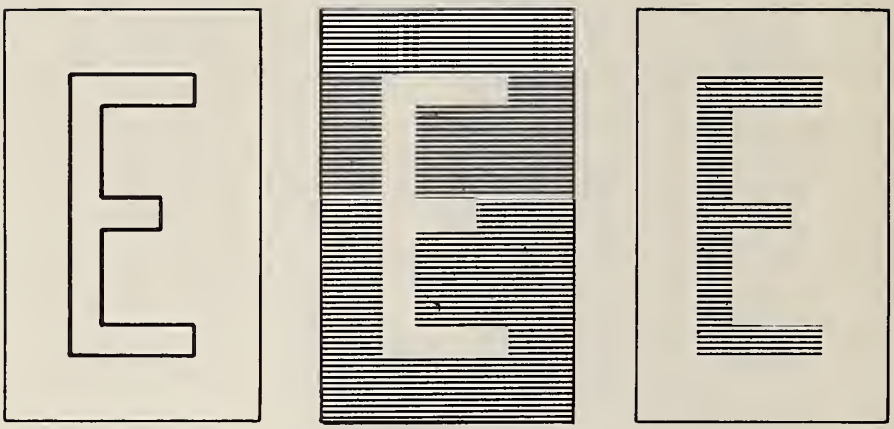
A changeover to optical character reading techniques is now under way, largely because of the interests of the Post Office as well as of a number of German banks in the increased use of high speed printing equipment. The Telefunken OCR-A character set model was demonstrated at the Hannover fair and has been undergoing further tests. Because this is a prototype, a relatively limited image area cannot adequately take line skew, so that higher reject rates are encountered than in the CMC-7 magnetic font, 16-pocket machines now in production. However, a new optical reading head has been designed to handle vertical misregistration or skew up to a space of three normal lines.

In future developments, Telefunken intends to explore the automatic recognition of OCR-B characters because of greater public receptivity to this font, even though preliminary estimates indicate that costs may be three or four times as high as for OCR-A recognition electronics. The new design, being developed under the direction of Dr. J. Braunbeck, includes two feeders, one continuous and one on demand, with speeds of 60,000 six-inch documents per hour. The new optical reading head is physically small, utilizing special photodiodes developed as a by-product of Telefunken's work on solar cells. A new reading station has been designed with two drums and with only a 60 mm distance between drum and reading head. (The "hurry up and wait" requirements for on-line operation dictate a minimal delay in response to computer demand.) Illumination of the image area is in the infrared range to minimize effects of off-color papers, overstamping, and the like. Iodine lamps are used, giving very intense light, and thus saving amplification requirements at the reading head.

The photodiode reading head provides 40 channels for a semiparallel scan, allowing three lines of 13 character-scan units each. The head is movable, so that a line can be read anywhere on the document page. (Dr. Braunbeck indicated, however, that there is probably very little demand in Germany for full-page reading machines). Strokes are first thinned, and stroke analysis logic is then applied. Further system details include the use of photocells to check the whereabouts of documents in the paper-handling chain, so that, for example, an extra 'long' document occurring when two items have not been fully separated results in an automatic document reject operation.

It is to be noted that the computer control required for account lookup on tape also enables making multiple analyses of test data, e.g., of the relative frequencies of rejects by individual character. Braunbeck's previous work has included the development of a polarized optical matching system, for which he holds a German patent (Fig. 4) [7]. He discussed briefly a Telefunken display console for air traffic control applications using a (Telefunken) TR-4 computer. Drawings can be made on the scope, alphabetic labels inserted, and the like. A manual, moving-ball control permits moving images and scan areas multidirectionally about on the display scope.

As a complement to the optical character recognition work at Telefunken, Konstanz, a group headed by Dr. Meyer-Brötz at the Telefunken laboratories in Ulm is investigating the use of resistor correlation networks with suitable masks (criterially weighted templates) for the recognition logic. Research investigations into problems of the desirable number of criterial points, centering, and noisy characters are in process. Attempts are being made to improve noisy characters before quantization. Different algorithms are being explored to determine discriminant weights in order to minimize the requirements for the recognition matrix, where the values for the resistors are proportional to the probability of the $i^{\underline{th}}$ point being black if the $j^{\underline{th}}$ character has occurred.



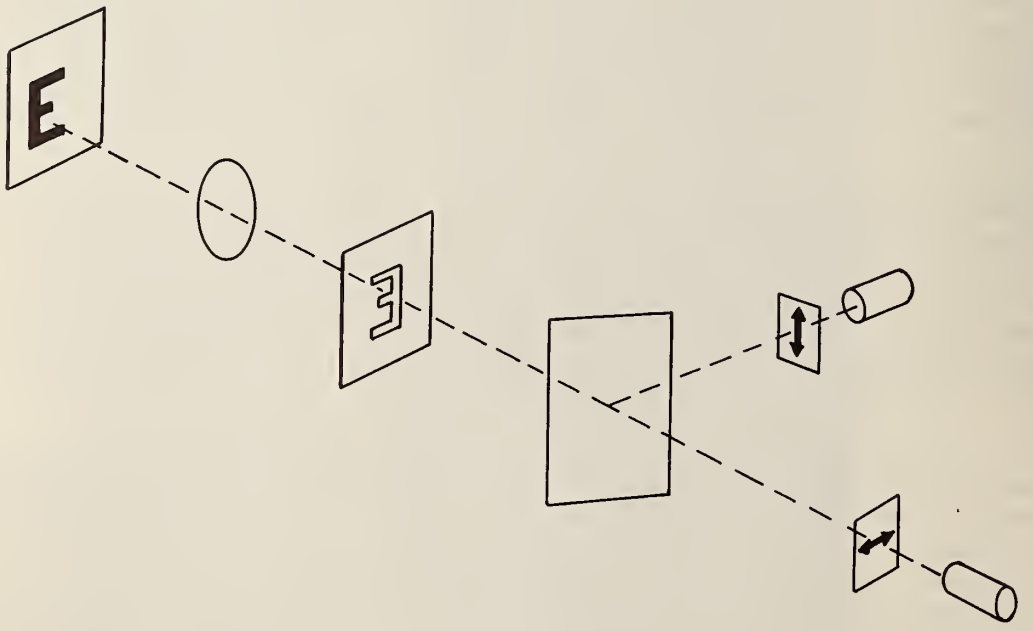


Figure 4. Detail from Braunbeck Optical Matching System Patent

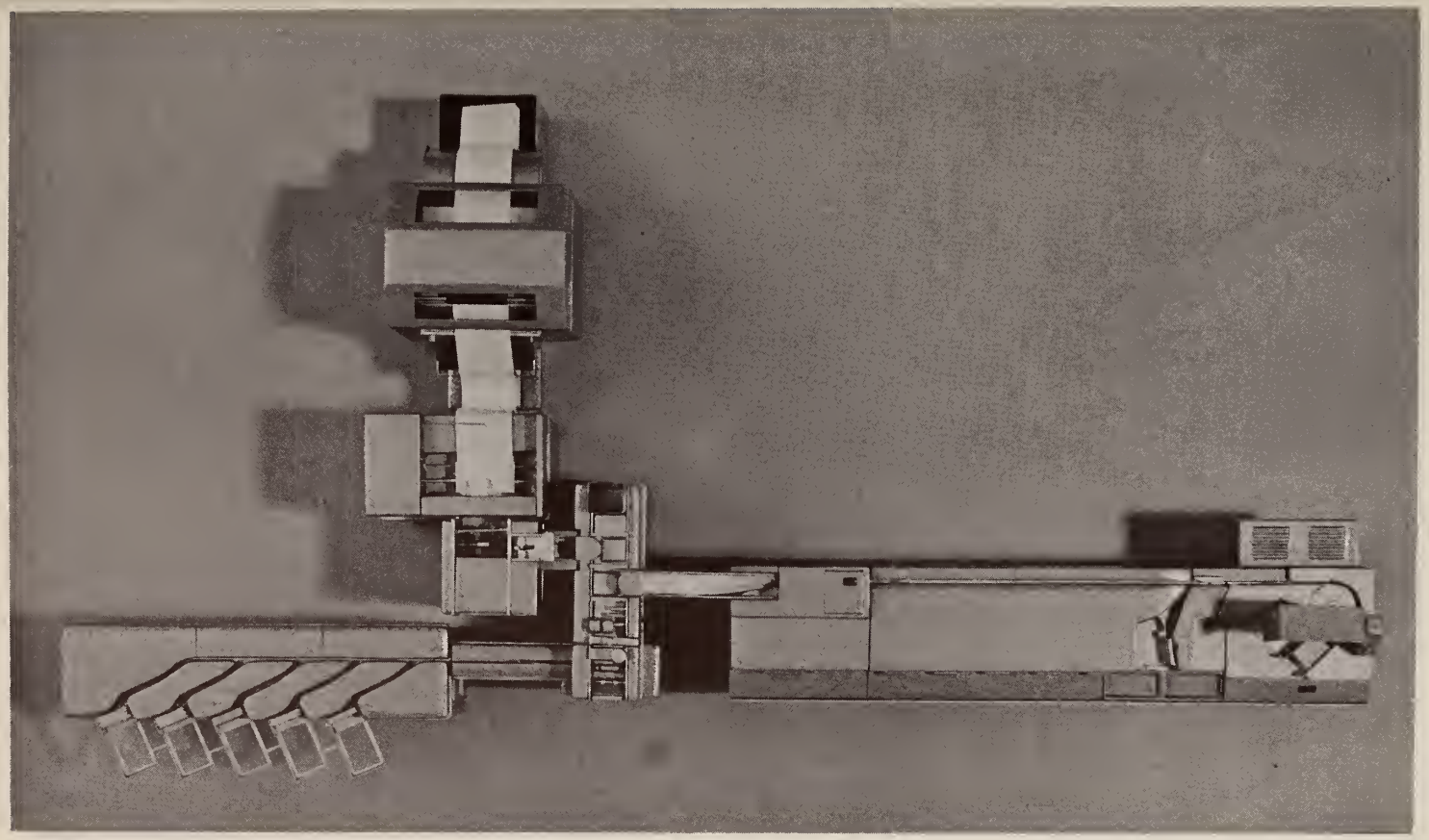


Figure 5. Standard Elektrik Lorenz ODS-2 Machine
(Photograph courtesy of Standard Elektrik Lorenz AG)

Possibilities for the use of mathematical regression analysis are being studied. First, the P_{ij} are quantized to + 1, 0, and - 1. A stepwise regression algorithm is then sought that develops the statistical dependence or independence of neighboring points, so that rows and columns of the matrix can be eliminated to get rid of linearly dependent elements. Meyer-Brötz is investigating the possibilities of finding functions that are non-linear and higher order correlations in order to identify and eliminate expressions whose factor is zero (hopefully, most of them). Simulations with a flying spot scanner and the Telefunken TR-10 computer have begun. The scanner provides a variable window for scanning, the smallest size being 3.75mm for typical typewritten characters and the largest cncompassing the entire page. Resolution is variable, from 8-10 points to 25,000 picture elements (100-micron-size).

At the central laboratories of Standard Elektrik Lorenz AG, Stuttgart, current SEL developments and interests in optical character recognition were discussed with Dr. W. Dietrich, who has been in charge of this work since 1957 [6], [8]. First, the ODS-2 machine (Fig. 5) designed to read the stylized numeral set of the OCR-A font is on the market. (The earlier CZ-13 font, developed at SEL, is very similar to OCR-A, and thus required relatively little logic modification, primarily changes to detect some horizontal stroke information, the CZ-13 logic having used only vertical stroke information.) Different configurations are available for differing customer requirements. Three machines with OCR-A logic are in experimental operation at the German Post Office, Nürnberg, (Fig. 6) and the Quelle mail-order house (the largest P.O. customer with 30,000 packages per day mailings), also at Nürnberg, is testing a system whereby it will imprint its own data, and the Post Office will add its information on another line, for documents of varying size and color backgrounds. The principal customers, aside from the Postal Service, are financial institutions interested in document sorting and account processing, and computer manufacturers (more than one in the United States, and at least four in Western Europe), who plan to use the machine for direct computer input.

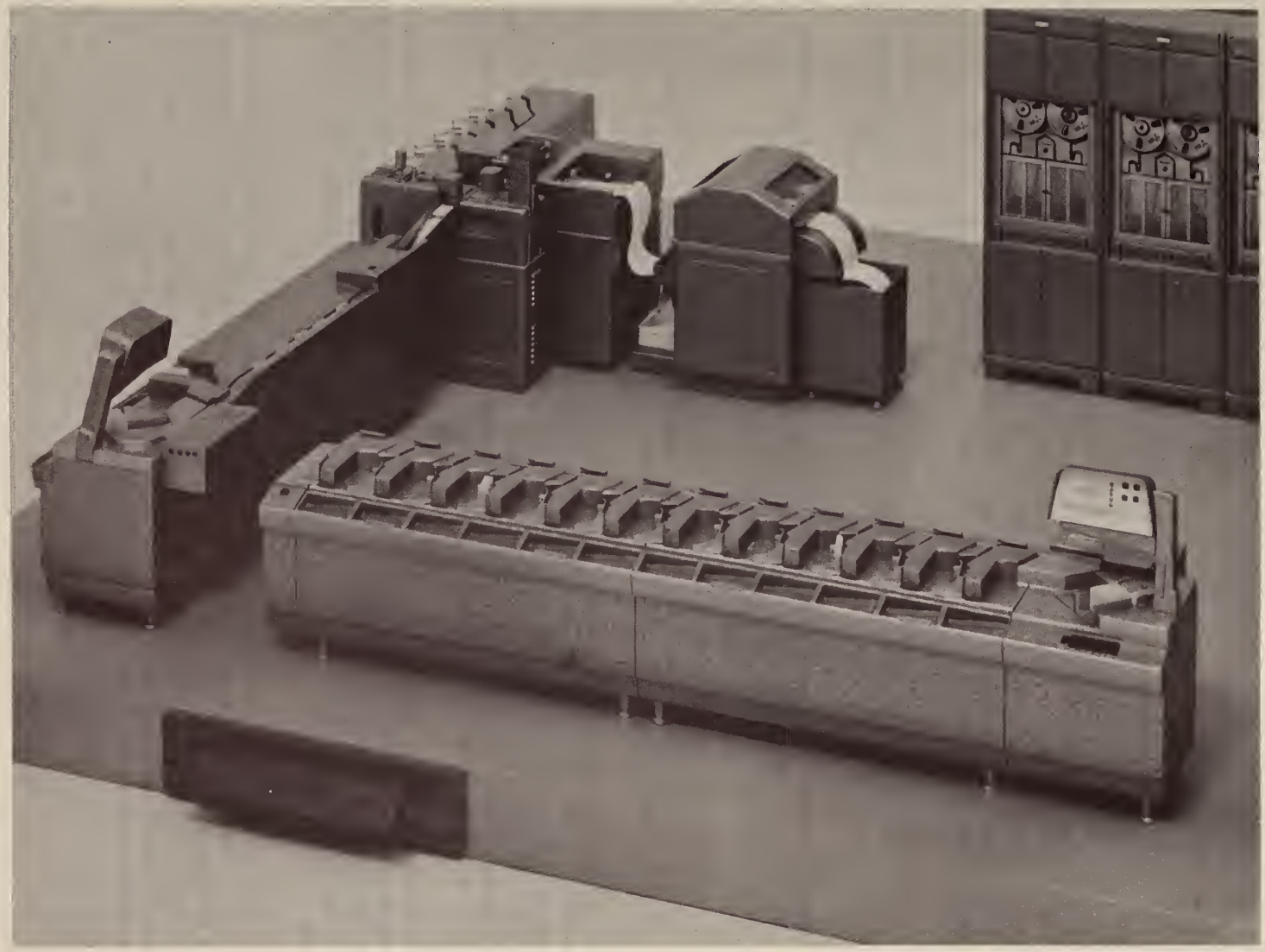


Figure 6. Experimental SEL Installation, Nürnberg
(Photograph courtesy of Standard Elektrik Lorenz AG)

The ODS-2 machine is designed with emphasis on the simplified, stylized font for maximum accuracy in the banking and financial applications. (It is of interest to note that Dietrich successfully demonstrated, in 1960, a reader for IBM pica typewritten numerics: he remarks that the new machines for the highly stylized fonts CZ-13 or OCR-A are some 30-40 percent less complex in terms of hardware and that they are faster as well as more accurate.) ODS-2 models sell for approximately $40, 000 to $50, 000, with 7 to 10 documents per second as typical speed performance.

In the machine, different scanning heights up to three normal character heights are provided for compensation of line misalignments. Basically a two-line reader, this machine can read lines anywhere on the document by the use of re-positioning mechanisms. Reading machines available at 1, 500 character/second can also be adapted to faster sorting machines, up to 3, 000 character/second. Infrared-range scanning is used to overcome most of the difficulties encountered with multi-colored overstampings, smudges and dirt, variable-color paper background, ball point pen super-impositions, and the like. Output from UNIVAC 1004 and IBM 1403 printers can be read.

Present developments are being directed toward a machine capable of reading the complete OCR-A character set. In order to deal with the comparatively small differences between some of the alphabetic characters, the following techniques are being investigated:

1. The character is projected onto a matrix of photodiodes and the analog values (rather than quantized digital signals of black-white distribution) are simplified, processed, and evaluated by resistor matrices.

2. In the case of critical character pairs, where the output differences are too small for reliable discrimination, additional recognition circuits are employed to differentiate between the two members of the particular pair.

As in the case of the numeric reader, low overall cost will be a dominant objective in the design of the alphanumeric machine. An operational model is expected to be ready in the near future. It is to be noted that SEL, unlike ICT and Telefunken, is committed to the highly stylized OCR-A, rather than the more conventional-appearing OCR-B font. The reasons given are primarily those involving the greater accuracy achieved because of the stylization to maximize difference between characters.

Dietrich has also made some investigations of possibilities for recognizing constrained handwritten numerals, to be inscribed in preprinted boxes with two dots, in a manner similar to the IBM and Dimond techniques. Dietrich suggests that multifont readers will not be of much utility in Europe, in part because of the very multiplicity of fonts that are widely used.

At Siemens & Halske, Munich, Dr. R. Jurk is continuing that company's OCR development work. Siemens is definitely interested in multifont recognition problems and has two major projects in this area: first, the recognition of postal area codes printed in different type-styles on different sizes, colors, and qualities of paper and, secondly, the recognition of microfilmed meterface readings [9], [10]. They consider that the higher equipment costs typically involved will be justified in terms of subsequent savings of later modification costs and in terms of flexibility to meet requirements of new areas of application at any time.

A first experimental postal code reader was demonstrated by Siemens as early as 1961. In the German system, the postal area code typically begins the second line of the address and is a 4-digit number preceding the name of the city or town. Approximately 90 percent of German mail is so encoded, with 70 percent of this being machine prepared. A prototype system for reading, sorting and distribution was demonstrated by Siemens at the International Traffic Fair, Munich, 1965, and advanced models are undergoing field tests by the German Post Office at Pforzheim. The current machines are capable of processing 18, 000 letters per hour and it is anticipated that this speed can be doubled by improvements in the mechanical transport system. Other adjustments will be made as indicated by practical requirements of observed field operations, including paper feed and maintenance service requirements.

Because of the Siemens experience to date, G. Gattner has stressed the importance of paper quality, paper dimensions, and effects on format control in the design of recognition systems. He suggests further that in the evaluation of character recognition techniques, a figure of merit should be obtained in terms of three factors: the size of the character vocabulary, the number of fonts that can be handled, and the print quality that can be tolerated.

The second OCR program at Siemens, that of meter-face readings, presents even more difficult challenges than those of the postal system. In the German telephone system, meter counters are set up in blocks of hundreds of subscribers. Periodically, the faces of the meter blocks are photographed onto 35mm microfilm, with control data added by teletypewriter entries. There are several millions of these meters in use, some of them over twenty years old, with a variety of type faces, and with the digits often misaligned and obscured by dirt on the glass windows protecting the meter faces. Moreover, since line width on the film is only a few mils, dust on the film itself presents additional problems. Because of the photographic reduction, also, it was found that the required scanning resolution of 3, 000, 000 points could not be accomplished with available CRT techniques. Instead, a mechanical system has been adopted, using a circle of mirrors and a vertical disc scanner. The disc itself is of plastic film, stabilized by air cushion, onto which the very fine, exactly positioned slits have been photographed. Characters not recognized in the system must have the location accurately identified for subsequent manual investigation.

In general, the recognition philosophy employed in both Siemens programs is that of the extraction of special discriminating features from character shapes, largely independent of character size, location in the input image field, and, to some extent, orientation (Fig. 7). Noise reduction is first applied to eliminate isolated noise and to smooth ragged edges (Fig. 8). As early as 1958, a computer program involving the application of majority logic to a small, sliding sub-matrix was developed. The features (or "form elements") used include determinations of convergence or divergence of character strokes (i.e., whether adjacent lines join or separate), and the detection of horizontal, vertical and diagonal strokes as well as arcs in various arrangements.

In the case of the postal code reading application, a two-step recognition process is required, so that, by detection of the structure of the address, the position of the code numerals may be determined. The Automatic Data Processing Division at Siemens is also interested in character recognition developments, particularly from the standpoint of direct computer input. A Siemens recognition technique for journal tape reading applications also involves quantization of 500 input emage elements and line-following to detect the "form elements". [9]

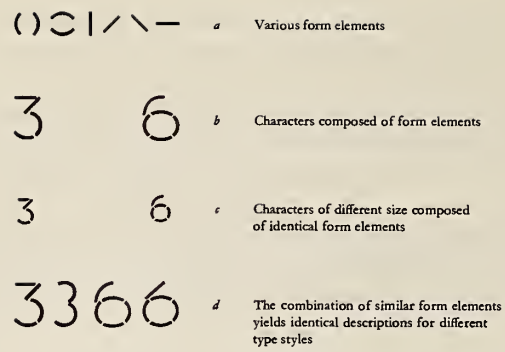


Figure 7. Differently Styled Characters Composed of Identical Form Elements

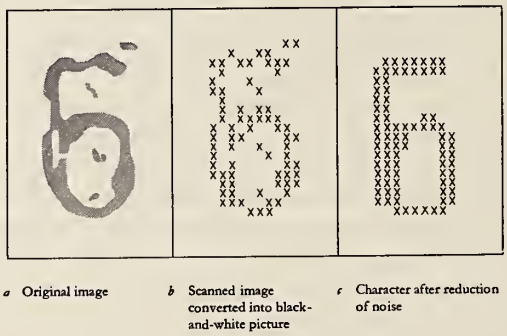


Figure 8. Character Image Enhancement by Reduction of Noise

Character and pattern recognition projects at the Technische Hochschule, Karlsruhe, under the general direction of Prof. Dr. -Ing. K. Steinbuch, include on-going research in the areas of multifont and handwritten character recognition and in the identification of signatures. For this purpose, an off-line input device has been developed that permits either conversion of light-pen input or of 24 x 36 mm film images to magnetic tape. This is operable in either a 64 x 64 or a 32 x 128 image-element mode. In particular, "a character recognition system for hand-printed numerals has been realized" [11]. Other investigations include work similar to that of Eden and Halle with respect to a limited repertoire of handwritten words, such as the words used in ALGOL programming. Simple shape-element features are extracted for recognition purposes. In the case of signature recognition, the account number is known and the problem is to determine identity for validation purposes.

Also under investigation at Karlsruhe are techniques for the automatic classification of handwritten characters, involving consideration of various methods of determining the discriminating criteria as derived from large samples of characters. Methods similar to those suggested by Kamentsky and by Vossler, and Simplex methods, are also being tried.

It is planned, when available, to use the integrated scanner-computer system to simulate linear classifying networks and to investigate automatic classification techniques under feedback control.

2.3 Italy

At Olivetti, Milan, Ing. A.G. Aghib of the Corporate Product Planning Department reported that, as of late 1966, Olivetti is concerned from a marketing point of view only with the printing aspects of OCR applications. Specifically, Olivetti offers typewriters equipped with either OCR-A or OCR-B fonts, and they are introducing these fonts for adding machines, accounting, and bookkeeping equipment.

However, Olivetti is continuing research on both magnetic and optical character recognition techniques. Several OCR readers are in the prototype development stage. The basic recognition concept involves a multi-topologic system for the detection of line directions, connections, intersections and relative positions of intersection as applied not only to the A and B fonts but to handwritten numerals as well. A distinctive feature of the Olivetti approach (i.e., in contrast to multifont techniques involving shifting from one font to another) is claimed to be the capability of handling a family of fonts so that the system can recognize characters varying as to pitch, size, stroke widths, and edge tolerances. With the kind of font family so far investigated, it is claimed that between 40 and 60 percent of characters used on typewriters in the United States can be accommodated. Another distinctive feature in the Olivetti prototype equipment is the use of ultra-violet illumination for reading. Aghib stressed the difficulties, especially for European applications, of problems of poor paper quality.

There have been several projects relating to character and pattern recognition at the Centro Studi Calcolatrici Elettroniche, University of Pisa. The first of these was a graduate student thesis project intended to provide formal descriptions based upon automata theory models for the various reading systems described in the Proceedings of the 1962 OCR Symposium [12] --- that is, for each system or device, to define in a formal way the input space and the transformations required to produce the given finite set of outputs. Unfortunately, the thesis itself was not very successful, but the approach suggested is of definite interest. A second thesis project, by engineering student Romano Tonini, has been concerned with the problems of reading photographs of telephone metering equipment (a problem similar to that under investigation at Siemens and Halske). The use of a vidicon system was proposed, but no hardware was actually constructed. Studies were made of typical shapes and discriminating areas for three different types of meter counter fonts.

Some pattern-recognition-related work is also continuing at Pisa with respect to the semi-automatic analysis of bubble chamber tracks, whereby operators of six tract-follower consoles measure tracks and feed the data to the computer which in turn provides feedback controls as to probable mistakes or errors. A time-sharing mode for these operations has been simulated via a punched paper tape buffer on both the CEP (Calculator Electronic Pisa), a 1960 machine of their own design, and on an IBM 7090 system.

Pattern recognition activities at the University of Genoa and the Centro di Cibernetica e di Attività Linguistiche, Milan, will be noted elsewhere in this report. At the University of Naples, a pattern perception study involves use of a curve analyzer to detect maximum and minimum inflections of lines and the investigation of techniques for coding lines, line connections, corners, and other characteristic features into a discrete number of symbols.

2.4 U.S.S.R.

A limited invitational Symposium on Mechanized Abstracting and Indexing, under the co-sponsorship of Unesco's Department of Advancement of Science and VINITI, the All-Union Institute for Scientific and Technical Information, Academy of Sciences of the U.S.S.R., was held in Moscow September 28 - October 1, 1966. During the Symposium sessions, Dr. A.I. Mikhailov and other Soviet scientists stressed that the automatic reading of texts is necessary for advanced research in mechanized abstracting and indexing and is essential if practical solutions to these problems are to be found.

With respect to character recognition, Mikhailov reported that Soviet scientists have visited centers in the United States, the United Kingdom and elsewhere and have also conducted their own experiments. They can therefore say that in those cases where there is sufficiently good quality control of the input material they can solve the problem with respect to input for index preparation. They have not solved problems of input of graphs and chemical structures. Therefore, the problem they are now trying to solve is that of chemical formats. This problem has not yet been solved, despite rather extensive experiments. Nevertheless, Mikhailov said he was convinced that these efforts would provide some practical solutions in the not too distant future. He stressed the universality of the input problem, noting that while it is difficult to achieve progress in the field, one can expect that, increasingly, the gaps will become fewer.

Considerable emphasis was placed by Mrs. M.L. Avrukh, of VINITI's character recognition project, on the importance of providing special devices for direct input from documents to machines, including the need for research because large masses of information may be best processed by character and pattern recognition techniques. She reported briefly on VINITI's character recognition device, which is intended to read material for their own abstract and index journals (e.g., to receive typewritten abstracts from their outside contributors) and on studies of the different typewriter fonts that will need to be accommodated. She indicated that they can 'read-in' (i.e., scan) pictorial data, but that VINITI is not presently studying pictorial data recognition, although other Soviet projects may be working on this.

On Friday, September 30, 1966, the Symposium participants were given the opportunity to visit the Laboratory for Electromodelling, where an operational character reader was demonstrated. Mrs. Avrukh was the principal spokesman for the project staff. In analyses of the problems in reading technical journal printing, many fonts are used and they have found that 500 to 1,000 different characters and symbols are typical. For prospective multifont recognition purposes, algorithms have been developed and performance has been simulated on computer. The practical objective is to provide direct aid to VINITI's information services, especially for the issuance of abstract journals and indexes. They are obtaining statistics on typewriter fonts used by their contributing abstractors, with the aim that input of prespecified format and good quality would be fed on-line to computer for the preparation of KWIC-type indexes and the like.

In 1964, a reader device was constructed to recognize a restricted class of typewritten text. Over 2,000,000 characters had been read as of the date of demonstration. Information as to the identity of the characters recognized is fed to a Ural 4 computer for experimental preparation of author and title indexes. The typewriter font used is large-sized and well-spaced (i.e., typed with a space between each character). The upper-case alphabetic and numeric characters used are somewhat stylized --- that is, characters that are normally wide have been narrowed, a small connecting bar has been added to the character "bl" (and even where this bar is defectively printed the spacing between the components is significantly less than that between characters), and punctuation marks have been accentuated.

5 ГI48 О ЦИФРОВОЙ ТЕХНИКЕ СЧИТЫВАНИЯИ
АНАЛИЗА КОНТУРОВ/ПОЛЯКОВ В.Г./
5 ГI5I РАСПОЗНАВАНИЕ РУКОПИСНЫХ ЗНАКОВ
С ПОМОЩЬЮ СЛЕДЯЩЕЙ РАЗВЕРТКИ/СЕМЕНОВСКИЙ/
8 ГI83 П ЧИТАЮЩЕЕ УСТРОЙСТВО/НАРИТА АКИРА
8 ГI76 К ВОПРОСУ ЭФФЕКТИВНОСТИ ПАРАМЕТРОВ
ОБРАЗОВ/ТАЛЬКСНИС Л.А./
8 ГI75 К ВОПРОСУ ОПОЗНАВАНИЯ ОБРАЗОВ МЕ
ТОДОМ МАСОК/БУЛОВАС В.В./

G 148 About Numerical Technical Computations and Analysis of Contours/Polyakov, V. G.

G 151 Recognition of Handwritten Symbols With the Aid of Curve Following/Semenovsky

G 183 Reading Device/Narita Akira

G 176 The Problems of Sensitivity to Parameters of Shapes/Talexnis, L. A.

G 175 The Question of Recognizing Shapes by a Method of Masks (?)/Bulovas, V. V.

Figure 9. Typed Characters Read by U.S.S.R. Reader

Figure 9 illustrates part of a typewritten page that was read during the demonstration, together with translations of the titles listed. (It will be noted that this is a case of well-shod cobbler's child!) With respect to quality, the following observations can be made:

> "Stroke thickness--Wide variation in stroke thickness from some characters to others and some variation within characters. Estimated stroke thickness ranged from .008" to .020".
>
> "Print quality--Ranged from fair to poor. I suspect that a fabric ribbon was used. Significant amounts of noise around character and in open areas within characters. Equally numerous voids in character strokes.
>
> My comments--A marked similarity between several sets of characters. With print quality as poor as the sample and the variations noted, it appears quite a task to differentiate between them.
>
> Punctuation--Sample only contained the . and /. Both marks are accented as far as size (or stroke thickness)." [13]

With respect to the recognition logic, it was not entirely clear from the Soviet scientists' presentation as to what features apply to more advanced models under development rather than to the present reader, but project personnel reported the development of digram and trigram statistics and provisions both for holding the recognition-decision for a given character in store until the succeeding character is "seen" and for use of "special areas to distinguish between patterns of near distance" based upon statistical analyses of pair-wise confusions.

The operating equipment, as demonstrated, appears to be neatly constructed, but with relatively obsolete technology in hardware by current U.S. standards. Document pickup and input is by precessing drum, scanned by flying spot scanner, and the signals are fed to a centering matrix and then to the recognition matrix.

The recognition logic apparently involves parallel comparisons of quantized input patterns against character reference patterns which seem to be of the weighted-area-correlation type, in the most general sense, and with "best-match" output. It was not clear, from an over-simplified flow-chart illustration, whether relative weightings are determined for all "coordinate positions" [i.e., an approach previously defined in the Soviet literature], or only for significant "information areas" previously determined by either manual or machine analyses of sample characters, which may involve feature extraction principles. A "peephole template" approach, in terms of criterial cells for specialized differentations between ambiguous-character subsets, would appear to indicate the latter emphasis.

Performance of this prototype VINITI reader system is quite slow. A typical page of the special-font input is read at the rate of about one page in two minutes. The input pattern is quantized to a 32 x 32 matrix, of which only about 300 cells are used for recognition purposes, a 24 x 24 matrix having been used for computer generation of the reference patterns for this particular font. The present error rate was reported to be one per thousand.

A new and more versatile character recognition system is in process of development at VINITI. Design specifications call for multifont capabilities, higher speeds, and improved reliability. Specifically, the improved reader should operate at a one megasecond cycle rate and should recognize at the rate of 200 characters per second.

Future plans also include research on pictorial data input. Up to the present, a scanning device has been developed for this purpose which, in these initial stages, provides 80,000 points per input item. However, processing programs must still be developed.

3. SPEECH ANALYSIS, SYNTHESIS, AND RECOGNITION

Potentially related to practical applications in the area of audio or voice inputs and outputs for non-numeric data processing systems are a variety of developments in speech analysis, speech synthesis, and speech recognition, including the special case of automatic speaker or "voice-print" identification. Projects in this area that were visited included the Royal Institute of Technology, the University of Bonn, the Telefunken Laboratories, and the Technische Hochschule, Karlsruhe. Projects at the Phonetics Laboratory, University College, London, were not visited because of scheduling difficulties, but progress reports were made available through the courtesy of Dr. D. B. Fry, the principal investigator [14], [15].

3.1 Royal Institute of Technology

At the Royal Institute of Technology, Stockholm, Sweden, there are various projects concerned with speech analysis and speech synthesis. Various members of the staff of the Speech Transmission Laboratory discussed problems of hardware developments for speech analysis, and research into effects upon normalization of speech in terms of the natural limitations of the speech organs.

The Speech Transmission Laboratory is headed by Dr. Gunnar Fant, whose interests in speech analysis date back at least as early as 1948. As outlined in a 1959 report, [16] these interests have included:

1. Study of the formal relations between the function of a band-pass filter and Fourier integral calculus, in terms of short-time frequency analysis by means of spectrographic devices.

2. Investigation of methods of evaluating transient response and temporal sampling characteristics of low-pass and band-pass filters.

3. Theoretical analyses of vocal tract transmission characteristics in terms of the theory of electrical transmission lines and of Laplace transforms.

4. Application of pole-zero specifications of transfer functions to problems of speech specification and of design of electrical analogs to vocal tract performance, and resonance analog synthesizers.

5. Experiments on tone-vowel associations.

6. Study of distributions of frequency-intensity formant data of speech as related to equal loudness contours and to perception criteria.

Several projects of the Speech Transmission Laboratory are concerned with aids for the deaf, both with training and with improved perception objectives. A. Risberg [17] demonstrated speech analysis hardware with rapid response displays of both spectrum and frequencies so that, for example, the totally deaf child can see the results of his own

attempts at speech as compared with the desired performance and can make instantaneous adjustments.

Apparently the study of tactual means for the reception and perception of speech dates back nearly 40 years. Thus the "Teletactor" was developed by Gault and Crane at the Bell Telephone Laboratories in 1928. This device operated by dividing the transmitted speech energy into five frequency regions and the amplified signal for each region was presented to one of the subject's fingers by means of a vibrator. Following the development of vocoder techniques, after Dudley (1936), a tactual output vocoder (FELIX) was built by Levine and others at M.I.T. in the period 1949-1951. The present tactual vocoder at the Royal Institute of Technology is designed along similar lines. [18]

It is to be noted that "vocoders" are devices "for compressing the bandwidth of speech signals in order to transmit them over channels of very limited capacity. The vocoder measures the speech power in a number of frequency bands and transmits these measures as signals over a set of narrow low-frequency channels... At the receiver the speech is reconstituted by modulating the spectrum of a broad band source in accordance with the frequency region and amplitude of each of the measure-signals derived from the original speech. Normally this reconstituted speech signal is presented acoustically for a listener. Alternatively, vocoded speech signals can be presented visually or tactually". [18]

The tactual vocoder at the Royal Institute of Technology first uses an amplifier to enhance the speech signal as it is received and then passes the signal through a differentiator to provide high-frequency emphasis. The signal is next divided by means of overlapping filters into ten channels of different center frequencies. The output signals from each channel are then rectified and smoothed to provide a control voltage and, in turn, each of the ten control voltages modulates the amplitude of a 300-cycle per second sinusoidal signal. When the varying 300-cycle signals have been amplified and adjusted for channel sensitivity, they are fed to ten bone-conduction transducers serving as vibrators that stimulate the tips of the user's fingers, proceeding from left to right across both hands in order from the lowest to the highest channel.

Discrimination and identification tests, for both normal and deaf subjects, have been conducted at R.I.T. using this experimental apparatus. Results showed some success in transmitting speech information through the skin by means of these relatively simple electronic transformations, but for some speech features only a limited level of performance was achieved. However, when combined with other means for information transmission, especially lip-reading definite performance gains were noted.

The investigators suggest that, in general, "when information is added to a limited sensory communication link, it can improve communication in two ways. First, if the added information conveys dimensions of the source code that are poorly transmitted by the existing sensory channels, then the total channel capacity is increased (for that particular source). Second, even if added information is partially or totally redundant and there is little increase in total capacity, the added redundancy will improve the resistance of the link to interference or distraction in one of the existing sensory systems. Thus, for example, in using the present tactual vocoder simultaneously with lip-reading, the vocoder provides new information for discriminating [the sounds] /m-b/ and the number of syllables, while providing redundant information for identifying some of the vowels. The new information extends the range of transmittable speech sounds and the redundant vowel information will tend to support vowel transmission..." [18]

Such findings may be of interest not only with reference to the development of improved aids to the deaf but also to situational or operational requirements where it is desirable to supplement or reinforce communication by presenting messages simultaneously in more than one sensory mode. Parenthetically, the question may be raised as to whether, if coupled to a suitable transducer of graphic patterns, such as the black-white distributions of alphanumeric characters, the tactual output signals might not be suitably encoded to provide an effective reading aid for the blind.

The problem of best design for a tactual code receiver involves that of developing codes having good compatibility with the reception characteristics of the skin, but not enough is as yet known about these characteristics. Questions are raised as to the extent of reception-perception limitations on serial combinations of tactual patterns and as to the extent to which these limitations may depend on factors such as practice, rate of transmission, type of tactual pattern, and inherent time-response characteristics of the sensory system used. The R.I.T. investigators conclude that although the ten finger-tip vibrators of the present model are convenient to use, the release of one or both hands would be desirable in most practical situations.

The work of J. Liljencrantz, U. Ringman, P. Tjernlund and others of the Vocoder Group, Speech Transmission Laboratory, R.I.T., is still primarily concerned with speech analysis and speech synthesis [19], [20], [21], [22], [23]. They hope to tackle problems of speech recognition, to which some attention has been given, at a later date, but so far there have been no practical accomplishments to report. Both a speech analyzer and a speech synthesizer are in operation and can be used together to improve the performance of the latter. A demonstration recording, in English, was quite impressive in showing adaptive improvement of the synthesizer output by the time that a third repetition of a sentence had been made by the speaker, whether male or female.

The speech analyzer presently in operation is a band-pass filter bank speech spectrum analyzer, with 51 channels that can be set in various combinations of center frequencies and different band widths. Input signals are digitalized after filtering in order to allow synchronous sampling of all channels. Output may be presented visually or recorded on magnetic tape in digital form, and it may be reconverted to analog form for display via CRT scope or on an X-Y plotter. Investigations can be made of effects of different frequency transpositions upon input signals, changes in the time code system, and the like, and the analyzer results can be compared with those obtained from the speech synthesizer.

Development of a new spectrum analyzer is well underway. It is being designed for on-line operation with a CDV 1700 computer, although it will be also operable off-line for purposes of monitoring, display, and recording. (It is to be noted that the 51 channels of the filter bank can serve not only as separate converters but also as intermediate storage devices in the case of delayed access to the computer when the latter is running in a time-sharing mode). The new analyzer is designed for systematic varying of formant and anti-formant information, e.g., to take samples of natural speech, to extract parameters, and to try to replicate them with the speech synthesizer. At present, these procedures are being simulated with a manually-controlled manipulator. (John Holm, of the British General Post Office, has been participating in this project as a guest worker).

3.2 University of Bonn

At the Institut für Phonetik und Kommunikationsfordung, University of Bonn, Federal Republic of Germany, work in speech analysis and recognition, and speaker identification, is under the general direction of Dr. Gerold Ungeheuer. Papers on topics of interest have been presented, for example, at the 5th International Congress on Acoustics [24], [25], [26], [27], [28], and in various progress reports of the Institute [29], [30], [31], [32].

H.G. Tillman is interested particularly in the development of a general model of speech communication from the point of view of linguistic structures and with emphasis on Ungeheuer's distinctions as between communicative factors. Under Ungeheuer's general direction (and, in part, active participation) hardware developments for speech analysis and recognition at this Institute include phase spectrum analyses and derivation of autocorrelation functions of fricatives in the initial position and equipment to measure width of intervals between zero crossings for clipped speech. There is a system of Ungeheuer's design that separates high and low frequency components, integrates filter outputs every three seconds, and provides output results to an X-Y oscillograph. In addition, a prototype word recognizer, DAWID, was begun in 1962 and was demonstrated at the ICA Congress in Belgium in 1965.

Ungeheuer's system provides the capability for photographing the different domains for movements (on the X-Y display, see Fig. 10) of coordinate positions specific for different speakers and of the characteristically different vowel triangles formed by the first and second formants. Further processes involve separations of the first and second formants by density distributions of zero crossings, and determinations of the number of zero crossings per 3-second intervals, giving other plottings. Using such data, it is possible to identify speakers independently of the text uttered, and significant differences between professional imitators and the public figures whose voices they are imitating can be shown.

A related problem is that of speaker identification when the speakers are uttering certain pre-selected code words. Tillman has analyzed 100 to 200 classes of sonograms of a particular word as uttered by 10 different speakers in order to detect differential features, particularly those of a general nature, extractable as "diffemes". This has been done by manual inspection and analysis to date, but it is planned to use a computer for further analyses.

In the word classifier developments, the objectives are to discover criterial features that are not phonemic segments but that will successively discriminate more and more different classes as additional features are considered. A speech recognition device utilizing these principles was constructed for twenty classes consisting of the ten numbers and ten selected words, in Italian, for the Euratom Center at Italy, Ispra.

3.3 Telefunken Laboratories

At the Telefunken laboratories, Ulm, West Germany, Dr. H. Kusch is working on speech synthesis, speech recognition, speaker identification, and general statistical studies of speech. Kusch's approach is based first of all on the assumption that formants (characteristic frequencies) do not provide sufficiently reliable information for speech recognition because of frequency displacements of sounds from one speaker to another. He states that: "Account has to be taken not only of the sounds themselves, but also of the sound transitions between consonants and vowels." [33] Kusch is concerned in particular with the derivation of characteristics from both the fine and the coarse structures of the continuous speech signal.

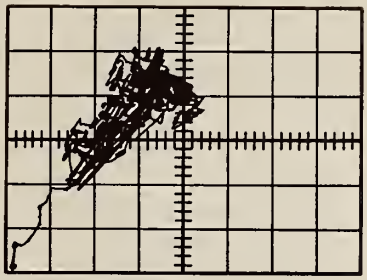


Abb.2: Sprecher S

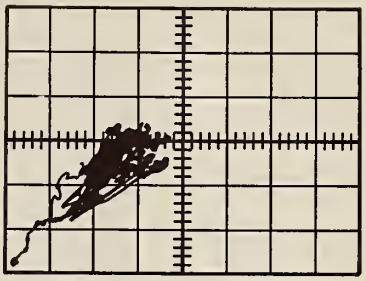


Abb.3: Sprecher Kr

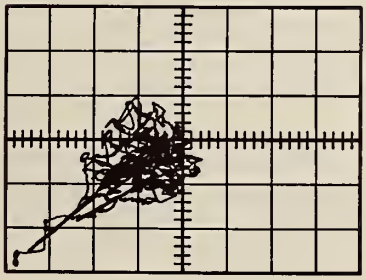


Abb.4: Sprecher R

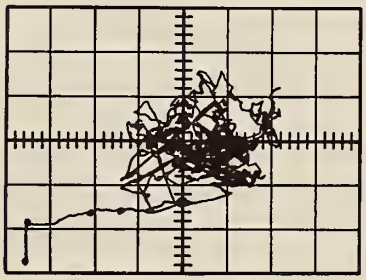


Abb.5 Sprecher U

Figure 10. Patterns for Different Speakers

For the recognition of spoken numerals (where, in German, a number of different pronunciations are found, e.g., for "sieben", such variants as "sieben", "zieben", "siben", "Ziben", "Siebn", "ziebn", "siem", "ziem", or "siebene"), Kusch first isolates an α-vibration and a β-vibration component of the spoken digit, using an analyzer having a high-pass filter and a low-pass filter with relatively flat attenuation characteristics so that the cut-off frequencies are non-critical. The output of each filter is fed to a Schmitt trigger which, in effect, quantizes the amplitude vibrations as present or absent (large amplitude - "present" and small or null amplitude, "absent") within sequential time intervals. The signals are then fed to flip-flops, they are subsequently scanned by a clock pulse generator, and the resulting information is transferred to the first of two coding matrices, for the detection of sound groups specified by the binary formulas for presence of absence of high amplitudes at five distinct time intervals. The second coding matrix relates the detection of presence or absence of sound-group indicia to the identification of the 10 spoken numerals, in either German or English. A sample of the circuit logic is shown in Fig. 11. It is claimed that, for well pronounced digits and for small differences in sound intensity, the equipment gives excellent automatic recognition.

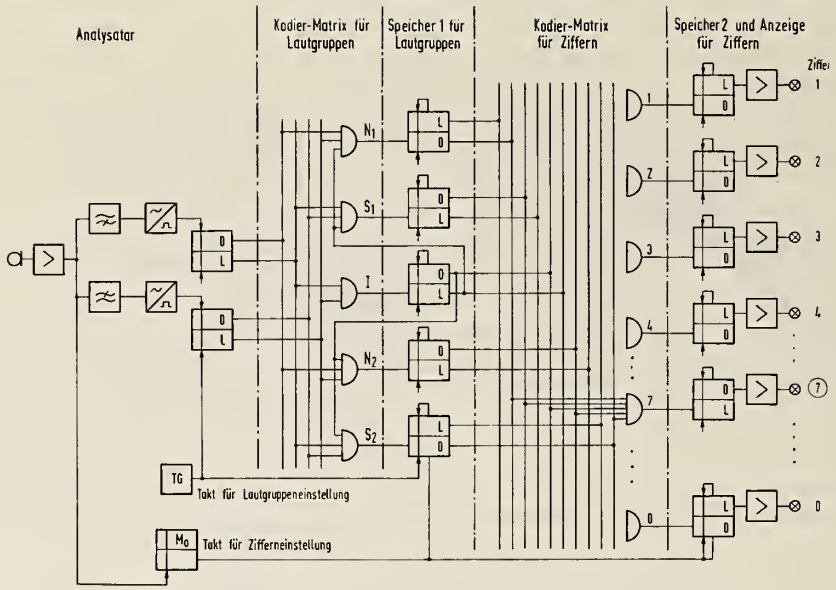


Figure 11. Circuit Logic for Identification of Spoken Numerals

However, in typical colloquial speech, many sounds may be pronounced indistinctly, too loudly, or too softly, often with large differences in sound intensity occurring within a single word. Therefore the β-component is divided into two parts with low and high sensitivities respectively. An extended recognition matrix for the three original sound groups and for a new group corresponding, in effect, to a pause (i.e., neither α- nor β-components are indicated as "present") can be developed, so that more accurate identifications may be obtained. In tests to date, for experiments with 37 male and 37 female speakers, an overall average recognition accuracy of 87 percent was obtained without adjustments of the equipment for the different speakers, and without compensation for wide variations of natural sound intensity and wide spreads in pitch. A second set of experiments involved normalization of the sound intensities to the same level and achieved improvement in average recognition accuracy to 93 percent correct automatic identifications.

3.4 Technische Hochschule, Karlsruhe

At Steinbuch's Institute for Information Processing and Information Transmission Technische Hochschule, Karlsruhe, Dipl. -Ing. von Keller is working on problems of speech recognition and has recently had his doctoral dissertation on these studies published. [34] He has explored several methods of speech analysis, including frequency analyses, autocorrelation, and zero crossings of the speech wave. He has constructed a correlator with variable filters to process speech sounds (especially sustained vowels) recorded on a tape loop.

It is planned for the future to track formants by autocorrelation functions for the eight German vowels, using real-time computer processing, but while these proposed methods have been manually tested, the computer has not yet been extensively used for the analysis of continuous speech. Von Keller points out that the machine measurement of formants automatically involves difficulties because formants tend to lie between lines of frequencies. He is therefore concerned with formant frequency determinations as being equivalent to determination of the first zero crossing of autocorrelation functions for the first two formants.

Other aspects of the zero crossing analysis approach include determinations of the time intervals between zero crossings on the speech wave curve and the number of such crossings that fill within certain time intervals. In particular, the investigator has tested a procedure, using 300μs. time intervals to determine the number of zero crossings that are of greater or lesser duration than 300 μs. to separate the different vowels. Using as criteria the mean lengths of distances as compared to certain pre-determined anticipated values, von Keller has already been able to separate five of the eight vowels.

4. OTHER TYPES OF PATTERN RECOGNITION

Nonnumeric data processing projects concerned with a variety of other types of pattern recognition than those of speech and alphanumeric characters are also to be noted. These include research and development activities involving adaptive and self-organizing systems and theoretical approaches to problems of artificial intelligence.

4.1 Other Projects at Karlsruhe

In addition to the character and speech recognition projects noted above, there is continuing work at the Technische Hochschule, Karlsruhe, on Steinbuch's general 'lernmatrix' approach to adaptive systems. [11] A non-binary (analog value) learning-matrix model was completed several years ago. Analog storage devices in the form of tape wound cores of small dimensions are used for the physical realization of the model. Stepwise setting of the cores in the matrix is achieved by coincidence of a pulse train of definite form and length with an RF rectangular wave. An additional RF current is used for non-destructive readout. Input patterns are applied either from a manual console or by photo-electric converters using a slide projector. H. J. Hönerloh has been comparing different methods of interconnection in 20 x 10 matrices of this type for the more effective simulation of "learning" processes.

Binary models of learning matrices are still under investigation from both hardware and theoretical points of view. One model of a binary predictor has been developed, using a learning matrix-dipole, which also employs tape wound cores as the adaptive elements. Experiments on over 100 binary sequences of 700 digits each led to correct prediction rates of 66 percent, on average, although the statistical properties of these sequences are such that at least 73 percent should be predictable in a-not-yet-realized system. H. M. Lipp is engaged in a project involving sequential binary signals to establish operating environments. Dr. E. Schmitt is investigating problems of optimization and control applications of the adaptive systems.

Dr. H. Kazmierczak reported that under a research subcontract for the German Ministry of Defense, investigations of automatic processing of pictorial data contained in aerial photographs are under way. The objectives include the location of streets and the recognition of objects such as vehicles and the determination of their coordinates. For this purpose, a flying spot scanner - computer system has been designed for implementation with a CDC 3300, and was scheduled for delivery in 1967. At present, research simulations are being carried out on a Telefunken TR-4 computer.

The scanner provides a variety of scanning modes and varying resolutions, later to be under programmed control, including line scans, point scans, control of a rotating spot of variable diameter around a point, and the capability of illuminating a line on the screen and rotating it for the detection of intersections with small, fine lines. Line scans may be selectively applied to areas enclosed by specified x-y coordinates. The present scanner provides a 256 x 256 scan-element area (8-bit coordinates), but it is hoped that this will be enlarged to accommodate 12-bit coordinate resolution later. In the aerial photograph problem, it is presently necessary to move the film mechanically for successive subsections of the photographic image and it is hoped that the developments currently planned will enable the future processing of the 9" x 9" original in its entirety.

In the detection procedure, the presence of parallel lines is first looked for at relatively low resolution, and then, within the indicated area, vehicles are looked for with high resolution scanning techniques. The scanning system provides pre-processing capabilities and it includes an analog-digital converter to process up to 64 levels of gray-scale information (6-bits). Pre-processing operations include evaluation of black-white density gradients, detection of contrast boundaries by detecting phase of signal, and determination of boundary directions. Six bits of recordable data are also available for encoding slope and directional information, obtained at a relatively low rate (125 kilocycles).

4.2 University of Genoa

At the Institute of Physics, University of Genoa, Italy, work has proceeded for some years on the application of adaptive and self-organizing principles to the design of physical models of pattern perceiving, classifying, and recognizing systems, under the general leadership of Prof. A. Gamba.

Essentially, Gamba's learning-recognition systems (the "PAPA" devices [35], [36]), are based on a criterial-crossings technique, where the interactions sought between input patterns and stored reference patterns are those of intersections of the input image with a pre-established intermediary pattern --- in this case, a configuration of random lines produced by a computer-based pseudo-random generator, and with associational- or probabilistic-inference weights built up from a training sequence of sample "characters" in terms of sub-units of the random configuration.

Specifically, the "features" or "properties" of the input image extracted include such possibilities as whether the total number of intersections are even or odd, the number of intersections in the first half of the image area as compared to those in the second half, whether more or fewer intersections occur to the left of the center of gravity of the input image than those occurring to the right, whether there are more or less than n intersections, and so on.

The acronym 'PAPA' stands, in Italian, for "Automatic Programmer and Analyser of Probabilities". [37] Palmieri and Sanna describe the original PAPA in part as follows: "Recognition is obtained through a statistical evaluation of identity or difference between the original information given to the machine by examples and the information derived from

the unknown pattern. The criteria for appraising a symbol identity are chosen at random without mutual correlation (although, of course, they remain unchanged during a given learning period and the subsequent recognizing). The recognizing process is then strictly probabilistic." [38]

An optical realization [39], [40] provided a filter for each of the two classes a and b to select only those optical fibers that give a preferential 'yes' answer for the appropriate class with weights proportional to their 'goodness' as discriminating criteria. Then, in later versions, it is noted that: "The 'random masks' of the old PAPA are now random lines from a cathode ray tube ... whose deflection plates are controlled by two magnetic tracks on which random noise has been previously recorded. We count the number of crossings between these lines and the pattern to be recognized --- every A-unit is 'on' or 'off' according to whether the number of intersections is greater or smaller than a preassigned value." [41] Fig. 12 illustrates one of several machine assemblies. Further, "PAPA No. 3" was designed to allow random units to be formed directly and continuously on a flying spot scanner driven from random noise generators." [42] Other pertinent references to the principles of operation of various "PAPA" devices include papers by Borsellino and Gamba [43], Palmiero [44], and Bertero [45].

In the presently demonstrable "PAPA" model, 4, 096 associational "cells" are used with 8-bit discriminatory capacity for each, so that 64, 000 bits of storage are required for this purpose and up to 8 categories or classes can be distinguished. It should be noted, of course, that size normalizations are not required in the system and that depending upon the training sequence, considerable tolerance to location in the input field and to rotation can be accommodated.

The present demonstration equipment and programming allow a 32-sample training sequence per recognizable output class. Output consists of the lighting up of an indicator signal for the machine-selected class and/or the print-out of the numerical values of the machine-calculated probabilities that a given input pattern belongs to class 1, 2, ... 8. The problems of squares vs. rectangles, with samples of different sizes, and of squares vs. circles of approximately equal area, were tried out during the demonstration. In the first case, results were very good when orientations close to true vertical or true horizontal were used for test inputs. However, with squares and rectangles close to the same total area (or with minimal discriminating line lengths), the results were often ambiguous. In the case of squares vs. circles of roughly equivalent areas, ambiguity and machine indecision were again to be noted. However, it is to be emphasized that only 32 training samples were used for both of these tests, and that many discriminatory problems may require several hundred training samples per class before correct distinctions can be achieved. In fact, in some potential applications studied, at least 500 samples per class (a not unreasonable number!) have been required.

Potential application investigations have included studies of the patterns of knight vs. rook moves in chess, and the question of whether the Pascal lines for random interconnections of lines drawn from 6 points on the circumference of a circle are vertical or horizontal (the latter point is a possible source of some of the confusion that has existed with respect to the PAPA technique. It is not essential to the methodology, but merely an example of discrimination by machine that would be relatively difficult for the human observer).

Perhaps more practically, the system has been applied to meteorological chart discriminations with respect to isobar or isotherm curve patterns that are characteristic, respectively, of good or bad succeeding weather in a given location. Here, given training sequences of 50 samples each for the two classes, the "PAPA" machine system has achieved 80 percent valid recognitions.

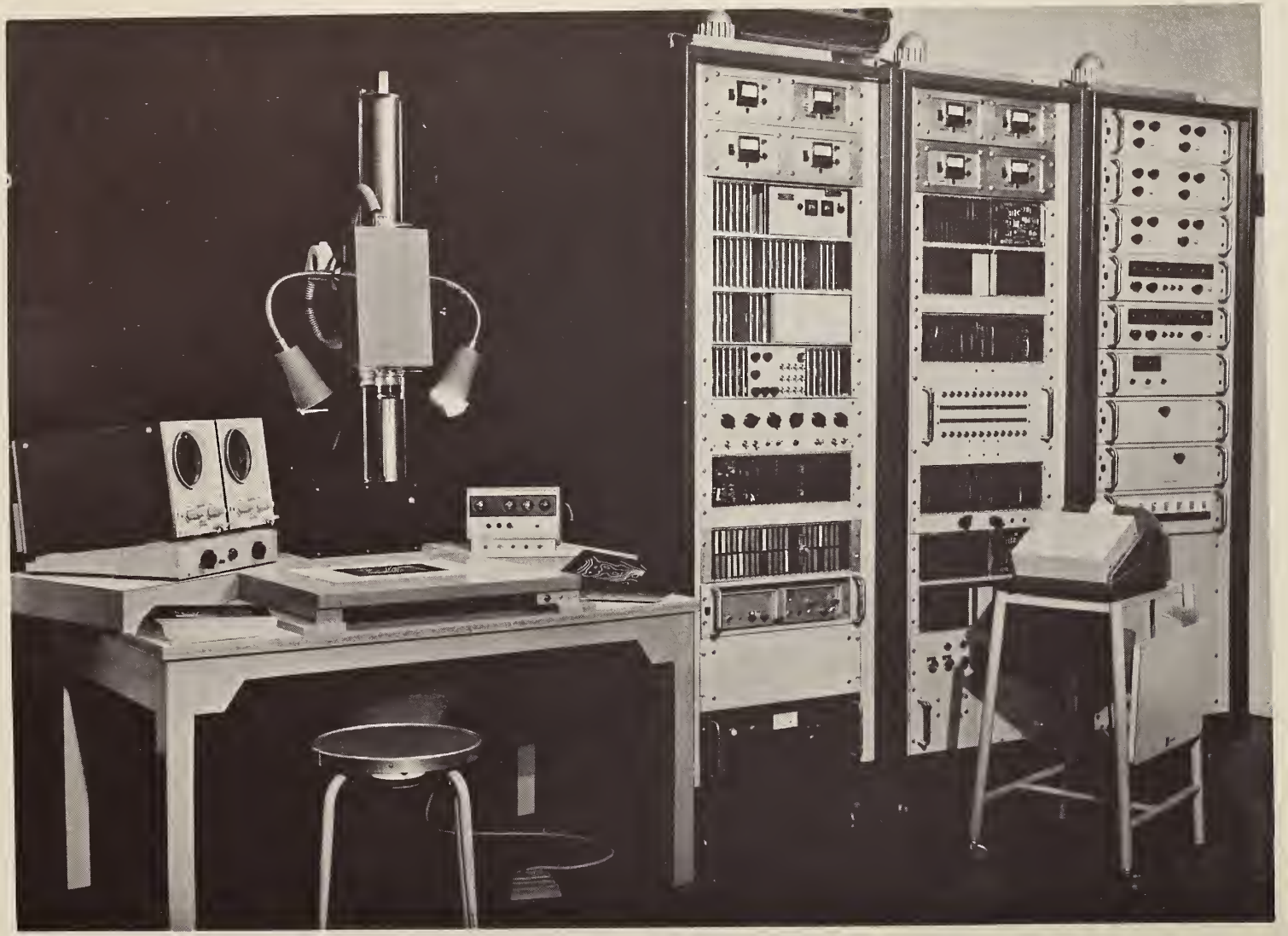


Figure 12. PAPA Machine Assembly
(Photograph courtesy of Università Di Genova)

To date, the same randomly-generated intersection-detection patterns have been used for the various tests. There is no reason, however, that other randomly-generated filtering patterns could not also be used, but this question has not been investigated in detail. On the other hand, at least preliminary consideration has been given to problems of speech recognition, without alteration of the basic approach.

The work of Borsellino and his colleagues in biophysics, also at the University of Genoa, is, at present, rather far removed from machine-processing considerations. However, they claim to have demonstrated not only learning but retention-of-learning (therefore, memory) in terms of transfer between larval and adult stages of the common mealworm, for T-maze training.

Borsellino and his associates are also interested in mechanisms of transfer functions of receptors, in the development of stochastic models showing the statistical properties of responses, and in problems of memory. Research in artificial memories is directed toward physical-chemical simulated systems, such as collagen "memories". In collaboration with Palmieri, they are studying complex feedback systems to explore possible mechanisms of axion action, connectivity of pulses, and currents through membranes.

4.3 Other Projects in Italy

In addition to the work of Gamba and his colleagues at Genoa, long-range investigations of problems of pattern perception and recognition and artificial intelligence continue at the Centro di Cibernetica e di Attivita' Linguistiche, University of Milan, under the direction of Prof. Silvio Ceccato [46], [47].

Ceccato's staff currently consists of seven people, with major interests in the study of psychological and mental activities. In particular, a series of investigations of movements of the eye is being carried out with a view to questions of attentional fragmentation, perception according to intentional instructions as against passive attitudes, vertical vs. horizontal fragmentations and differences in the field of vision as between different subjects. In addition, work is continuing on the construction of the perception-learning machine translation, the latter work emphasizing the determination of effective semantic classes.

Over the past fifteen or more years, the "Italian Operational School", led by Ceccato, has sought to structure epistemological phenomena in terms of actual operations of perception and apperception, with the expectation of developing machine-useful encodings. He and his associates have stressed the study of language and learning in terms of constitutive mental activities and more specifically in terms of ways of directing and using attention and of forming or perceiving relationships as plays of attention. Several hundred basic relationships can then be encoded as states and combinations of states of attention, maintenance of a state of attention as another is added to it, and memorizing a prior state while a new state or combination of states is taken up [48], [49].

In the area of linguistics, generally, Ceccato points out that: "One of the new and extraordinary chapters in operational linguistics is that which has to do with examining how much and what kind of culture we need to understand a text in a univocal way." [50]

Collaborative projects between the Centro Studi Calcolatrici Elettroniche, University of Pisa, and the Faculty of Medicine of that University are concerned with automation of the analysis of radiocardiograms. This involves a non-linear programming problem to determine transfer functions of interest, requiring detections of the first three moments, average dispersion, symmetry, as characteristic of desired curves. Similarly, there is a cooperative project with the Department of Neurophysiology to extract time sequences of nerve pulses and to determine coding in the optical system. For this purpose Prof. Marruzzi has designed a special-purpose analog-to-digital converter. A more generalized converter is under development and should be ready in about 6-8 months for a variety of biomedical investigations including EEG, EKG, and other problems involving pattern recognition with respect to clinical data.

4.4 National Physical Laboratory

In continuing work at the Autonomics Division, National Physical Laboratory, Teddington, Dr. Bryan Richmond is doing special programming for pattern recognition research such as the simulation on the KDF-9 computer of techniques to pick out local lines, to reduce bandwidth requirements for pictorial data processing, to detect contour line directions, and the like. An NCR Elliott 4100 computer is used for programming research studies as such, particularly for the development of programming systems to do binary operations conveniently, which conventional programming languages such as ALGOL do not provide. Facilities presently available include a display system to indicate visually, for example, the effects of the log probabilities by character for n times a function "g", so that one may determine desirable positive and negative weightings of character features and select functions suitable for linear decision-making.

Research into pictorial pattern processing at NPL, (e.g., fingerprints, chromosomes), takes advantage of boundary and contrast enhancement techniques developed there. Mr. Ralph Rengger discussed and demonstrated some of the pre-processing techniques in use. These techniques are based on Fourier transforms resulting from apertures placed in the plane of an imaging lens to suppress selected spatial frequency components. The basic pattern to be processed is in the form of a photo transparency, illuminated by a laser or other monochromatic light source whose collimated output is brought to a focus by a lens. The lens in turn re-images the input, via various Fourier plane and band limiting stops, to a processing-image plane such as a TV camera, whose video signals may then be fed to a pattern-processing or recognition system.

Demonstration photographs for output results for chromosome photographs were quite good, showing both outlining effects (elimination of areas of relatively constant gray-scale density by low-frequency-component suppression) and reduction of noise and fine granular structure (by suppression of high-frequency components).

John McDaniel, formerly associated with NPL machine translation projects (now inactive), discussed current interests in machine processing of palintypy recordings (equivalent to stenotyping) and question-answering or fact retrieval systems. With respect to the first area, he is concerned with a longest-match dictionary search procedure, using storage disks and the KDF-9, with the objective of producing hard copy from verbatim proceedings at the editing stage, e.g., from proceedings of Parliament or the law courts. In the second area, he is concerned both with investigations of potential current applications and with on-going basic research. Possibilities for current application include police records, staff personnel records, and telephone directories where the remotely accessed store would be organized by addresses rather than by names. Another possibility being studied is the simulation of situations such as air traffic control where natural language input statements would be transformed by automatic analysis into formal logical statements,

and logical answers would be obtained and transformed into the appropriate executive orders via a generative grammar.

Finally, at NPL, E.A. Newman discussed his continuing interests in voice recognition and in psychological research related generally to problems of pattern recognition. In particular, he is exploring such questions as whether the habituation effects to a stabilized image on the retina have equivalents on the acoustic side such that, by determination of the fragmentation as parts of the pattern are temporarily lost, components suitable for recognition may be identified. One type of pattern break-up that has been observed tends to support the phoneme separation principle, for example.

4.5 Sweden

At the Swedish Royal Institute of Technology, Stockholm, S. Karlsson of the Institute for Telecommunication Theory is engaged in the study of problems of pattern recognition with respect to electrocardiographic recordings. He and Dr. Arvidsson have developed a programming system for EKG measurements, using an approach somewhat similar to that of Pipberger at the U.S. Veteran's Administration. Karlsson is especially interested in the problems of finding useful clusters in the data. He starts with a procedure to formulate initial clusters such that a member of a cluster has its nearest neighbor in the same cluster (e.g., in the peaks of the recording), and repeats for the next nearest neighbors and so on. The purpose is to derive a measure for each pattern category by extracting features common to all members of the same cluster.

More specifically, the mean variance of each pair of items is determined, a quadratic measure is used to minimize error, and all items in the sample are categorized. The procedure is repeated, new weightings extracted, and a new iteration initiated. After five-to-seven categories have been obtained, they are evaluated to see if they are reasonable in the sense that the least distance between members and non-members is greater than the greatest distance between members. Consideration is also given to hyperplane boundaries. The procedure is presently limited to 1,000 or fewer items because of computational difficulties.

Karlsson is concerned about problems of extracting representative samples from a larger collection of data, but in fact he said that he has not as yet acquired a large enough sample of basic recordings. Results to date, based upon 50 samples each of normal subjects both before and after work and of normal subjects against subjects known to be suffering from coronary insufficiencies, show a 9 percent error by comparison with human judgment in the first case, and a 6 percent error in the second case.

5. LIBRARY AUTOMATION, MECHANIZED DOCUMENTATION, AND ISSR SYSTEMS

Some of the nonnumeric data processing projects visited during the European survey were specifically concerned with experiments, new developments, and practical operations in advanced techniques of library automation, mechanized documentation, and information storage, selection, and retrieval (ISSR) systems.

5.1 The "Bibliofoon" System

At Delft, the Netherlands, the University Library has come to be recognized as the central technical library and reference center for Dutch industry. (A similar arrangement prevails in Hannover, Federal Republic of Germany). Some 300 requests per day for service and loans are received from within the University, and some 500 requests are also received daily from industry or other outside organizations.

Mechanization in this Library is directed toward the location of books in, and the delivery of books from, the stacks. Since, to date, only one copy of each book is acquired by the library and because of the volume of demand, there is generally only a 50-50 chance that a given book will be "in" and available for inspection or loan. Therefore, to speed service to the client (either to deliver the requested book to him, or to notify him that it is not available), the "Bibliofoon" system has been installed. The 350, 000 items in the collection are stored in shelves, stacks, and block of stacks, arranged in order reflecting physical convenience (e.g., size of volume). The various card catalogs (author, keyword, and classified) are then relied on for the determination of the call number, which is the precise storage address. Thus the first three digits of this call number identify the block of stacks; the fourth, the stack within the block, the fifth, the shelf, and the sixth and seventh, the book's position on the shelf.

The client has available in the catalog file area, several dial telephones which he uses to dial the call number. On the appropriate floor of the stacks, a bell rings and an attendant is guided to the proper section of the stacks by lighted signals for block, corridor, and shelving, while another lighted indicator shows the last three digits of the call number. A central switching system takes care of changes in actual locations --- that is, books retain their call numbers even when moved to new locations but the lighted signal indicators will correspond to the new address.

When reaching a specific address, the attendant pushes a button to indicate whether the book is in fact in, or out. In the latter case, a message goes to the loan desk and the call number is typed out on a typewriter there. If in, the book is taken to a spiral plastic chute where it is dropped and subsequently moved by conveyor belt to the charge-out desk. (It is claimed that exhaustive testing and use of the chute system has shown that wear on the books is negligible, far less than would be suffered by one trip on the client's bicycle in a shower!) Every request, and whether the book was found to be in or out, is recorded on punched paper tape so that statistics useful for analysis of loan distribution and needs for additional copies can be accumulated.

Dr. J.W. Zwartzenberg, the Deputy Librarian at Delft, reports that the Bochum Library near Cologne is probably the most modern in Europe, with a punched card acquisition system, magnetic cards for both book and client identification, and automatic follow-up on the return of loans. Their author catalog is mechanized, and they are working on their subject classified file.

5.2 The EURATOM Center in Burssels

The Center for Information and Documentation (CID), EURATOM, is headed by Dr. Rudolph Brée, Director. The mechanized documentation service at this Center, using a computer-based thesaurus lookup and retrieval system, became operational in May 1966, on an IBM 360/40 installation. The collection consists of some 350, 000 items in the fields of nuclear science and technology and also in fields of peripheral interest such as the chemistry and metallurgy of reactor materials [51]. About 70, 000 new abstracts that are not covered by Nuclear Science Abstracts are received each year.

The problem of two or more abstracts for the same item appearing in different journals is attacked on initial input by manual assignment of a standardized identification code (author, year of publication, report number, etc.) and subsequently by computer checking for duplicates. Later, it is hoped that information will be extracted automatically from the magnetic tapes available from AEC's Division of Technical Information Extension, Oak Ridge, Tennessee.

Scanning and indexing services are provided by some 50 or more European scientists on a contract basis and a cooperative agreement between the CID and the U.S. Atomic Energy Commission has been in effect since 1964 to provide for descriptor assignments to the material covered in Nuclear Science Abstracts.

Under these circumstances of decentralized indexing, the problems of terminology control have been particularly important. Basically, the selection and retrieval system is of the coordinate indexing type, with noise from possible false coordinations controlled largely by indexing down to small sub-sections of the document as necessary. The EURATOM-Thesaurus that has been developed presently consists of 4,470 descriptors, of which 3,240 are names of specific nuclides or inorganic compounds. The thesaurus is considered as a checklist, and indexers are instructed to assign every pertinent descriptor applicable to the document.

The indexer may also assign specific terms in addition to the established descriptors if he feels that they provide additional information of value for search, such as names of alloys, reactors, projects, and the like. However, the extent to which usage of specific non-descriptors must be related to descriptor usage is carefully controlled, so that subsequent searches, whether generic or specific, will nevertheless retrieve the item. For example, if "Inconel alloys" is not a descriptor it may be used, but the indexer is instructed to assign also "nickel alloys", "cobalt alloys", and "chromium alloys". In subsequent computer processing, indexer compliance with these rules is checked and it is also possible to supply the appropriate cross-references automatically.

Key features of the system are computer printouts of the current thesaurus and dictionary of allowed additional terms, together with statistics of the most recent run and cumulative frequencies of usages. Other aids to the indexers are provided, especially "arrowgraphs" showing graphically hierarchical and other relationships both between descriptors and between terms (Fig. 13). A very high level of inter-indexer consistency is claimed. In a paper by L.N. Rolling [52] the following data are reported with respect to one test in which 35 indexers each indexed 20 physics abstracts and another test in which two teams of indexers analyzed a set of 2,331 abstracts chosen as an 0.7 percent sample of the collection: "The values of the relative and absolute, formal and conceptual consistency were determined for various subject categories in the nuclear field. An average conceptual indexing consistency of 90 percent was found... The values corresponding to the different subject categories were scattered between 78 and 98 percent; this reflects variations in (a) the background of the indexers; (b) the stability of the terminology; (c) the density of the thesaurus vocabulary." Rolling notes, however, that a nonspecialist in the subject matter "even with experience in documentation, will seldom attain as much as 70%" indexing consistency. There is, thus, at the Center for Information and Documentation, a strong emphasis on subject matter competence in the indexing.

The computer-compiled statistics of descriptor usage frequencies are also of importance in the searching operations. Queries from clients are first translated into the vocabulary of the system and are re-formulated into several levels of AND-OR-NOT specifications, that is: "tight", "loose", and "intermediate" requests with respect to selection thresholds for both high relevance and high recall. The cumulative usage frequency data is used, first, to determine the relative probable selectivity of groups of terms in the

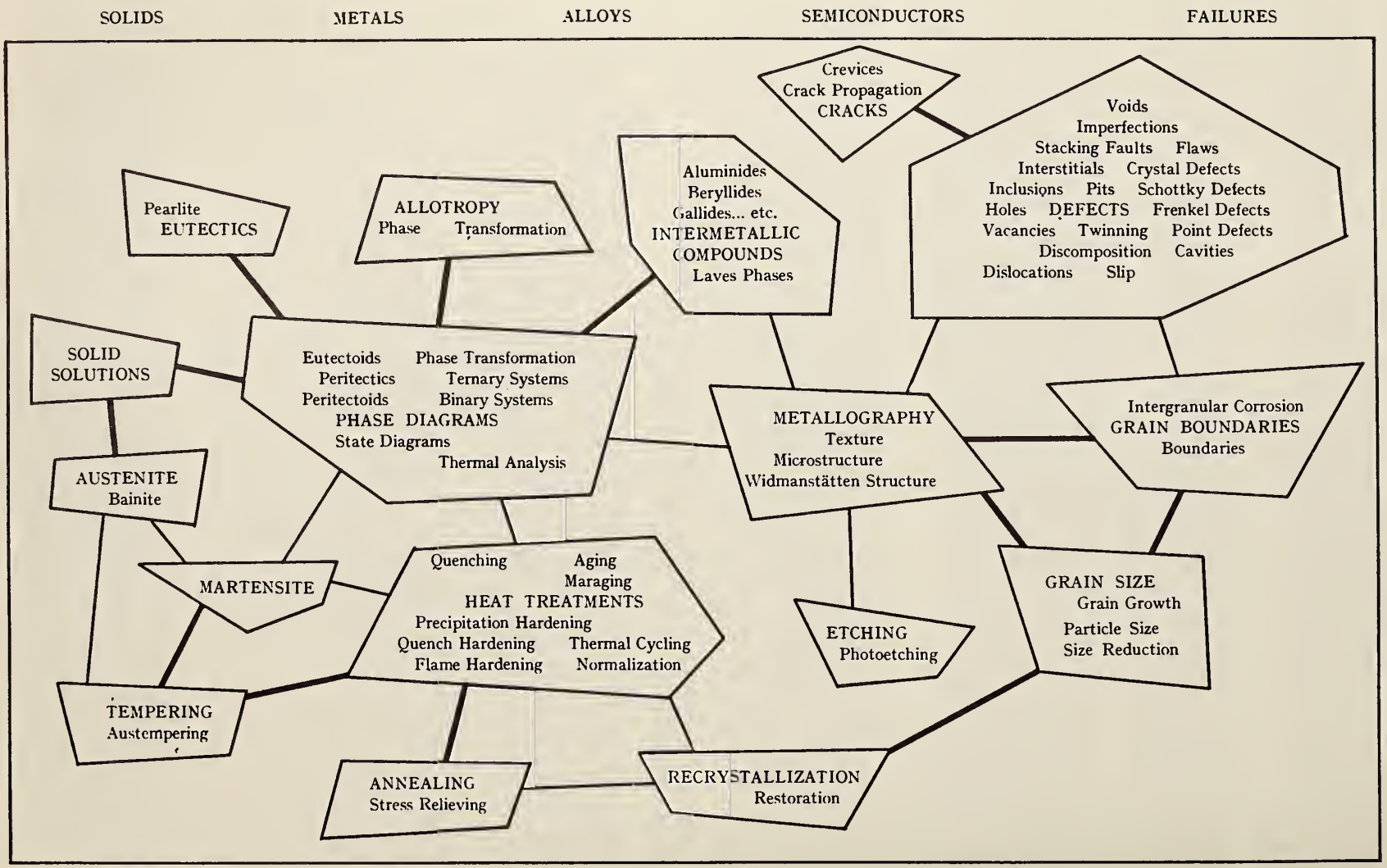


Figure 13. EURATOM Thesaurus Display of Relationships Between Indexing Terms

original query (in the process of setting appropriate search strategy, i.e., lowest frequency term search first) and, secondly, to provide an estimate in advance of the probable order of magnitude of the number of items likely to be retrieved by a particular query, so that the search prescription can be broadened or narrowed appropriately. The formula used to obtain the quantitative search-result estimate is:

$$R^n = \frac{f_1 f_2 \cdots f_n}{V^{n-1}} K_n$$

where: n is the number of descriptor groups (i.e., the term or terms separated from another term or terms by the AND operator);

f_i is the sum of the frequencies of usage of the descriptors in group $_i$;

V is the number of items in the collection;

k_n is the association factor.

It is noted that "the association factor owes its existence to the fact that only a minor part of all possible associations of n descriptors is meaningful; its value is determined statistically." (Rolling, [49]). It is probable that the association factor can be determined only by experience with actual queries. For example, in a test case, there were six items properly responsive to the query for the term "man" and the term "quantum mechanics".

Prior experience with the manual system has indicated a customer load of at least several hundred queries a year for in-house personnel. (Later it is planned to extend the services to others outside EURATOM.) In the three months between May and August, 1966, 44 in-house queries had been processed on the computer system. In addition, members of an advisory group were invited to pose questions to the system. A copy of the results in chart form is shown on page Qualifying comments regarding their statistical evaluation are as follows:

"When studying the table giving the retrieval results of the questions put by the members of the C.I.D. the following remarks have to be taken into account.

"The questions are "translated" into keywords in such a way that they are acceptable by the computer. In many cases this translation is done in such a way that one receives different levels of questioning. These levels are distinguished by the self explaining terms 'tight', 'medium', and 'loose'.

"In the table the number of sub-queries indicates whether one, two, or three levels were used. Nevertheless the screening was restricted to one level only.

"This was done here because the searches were mainly meant to be a demonstration of the way the retrieval will be done and not yet a performance of a smoothly runned-in [sic] information department.

"The reasons are amply explained by Mr. Brée during the meeting. It is further pointed out that not all material is already stored, which means that there might still be gaps in the information resulting from the search.

RETRIEVAL

Question formulated by	Subject	Number of sub-queries	Number of retrieval answers/subquery	Number of screened answers	Hits	Possible Hits	Relevance ratio (%)	Remarks
1. Dr. Kramers	High temperature measurement in neutron fields with the aid of thermoelements.	2	27/79	27	16	4	74	
2. Dr. Kramers	Neutron spectra of fast critical assemblies.	2	16/26	16	2	4	37	
3. Dr. Kramers	Neutron scattering by H2 0 ice and D2 0 ice at all temperatures.	2	18/20	18	11	3	77	
4. Dr. Kramers	Hydrides of Nb, Ta, Pa (Phase diagrams, structure at low H-concentration).	1	49	48	17	11	58	
5. Dr. Kramers	Multigroup Diffusion Codes (one-, two-dimensional).	1	335	335	39	-	(11)	screened for codes only
6. Mr. v.d. Laan	The applications in Europe of atomic energy for the desalination of sea water with special reference to the costs.	3	16/18/54	16	1	14	93	
7. Mr. Cacciapuoti	Irradiation de produits agricoles avec rayons gamma.	2	62/131	127	82	33	90	
8. Mr. Guilloux	Etat actuel du traitement par rayonnement du cancer du sein.	2	103/143	99	66	17	83	
9. Mr. Terzi	Architecture of structure of genetic inform. and variation of it under effects of nucl. radiat.	1	118	23	13	-	56	query not quite understood, answer only partly screened

"An answer is marked as 'possible hit': 1^{o} when the question itself is not absolutely clear and it is felt that the answer comes within the broadest meaning of the question and 2^{o} when, although the question is clearly formulated, it cannot be decided from the available abstract whether the answer is relevant or not.

"Finally it is underlined that in the operating phase the machine printout of keywords used will not be made available; the customer will receive copies of abstracts." (Verhoef, Aug. 22, 1966).

In addition to retrospective search services, a Selective Dissemination of Information (SDI) system is planned, and work plans for the next quarter by project are now being translated into the keyword system for use as project profiles.

The present staff at CID, Brussels, consists of 13 scientists, plus 4 more doing manual searches (especially for queries outside the scope of the computer-based system), with supporting clerical staff. As noted, outside contractors are used for literature scanning, indexing, and also for much of the coding. Where searches cannot be fully serviced from CID's collection, advantage is taken of cooperative exchange agreements with more than 120 other documentation centers.

5.3 Mechanized Documentation in Western Germany

The Institüt für Dokumentationswesen, in Frankfurt (Main), headed by Dr. M. Cremer, is financed jointly by the Federal Government and by the Community of West German States. It is formally organized as a part of the Max Planck Institute, but in effect it is a separate organization with its own board of trustees and advisory board. Its functions are similar to those of the Office of Science Information Services, National Science Foundation, in the United States, although on a somewhat smaller scale. The Institute has a staff of 60 people and an annual budget of 3,000,000 deutsches marks, (i.e., about $750,000). The Institute coordinates, promotes, and finances documentation projects on a national level. It provides seed money for new projects (typically, for the first two years' operations of a new activity), funds for training and the provision of special courses, and money for libraries and documentation centers to procure teletypewriter equipment prior to more comprehensive mechanization. About 100 different projects are currently being supported.

Special interests in the promotion of machine techniques for library use led, in 1964, to the establishment of a separate organization, the Zentralstelle für Maschinelle Dokumentation (ZMD), to provide a central computer service for documentation and library applications. The two institutions continue to collaborate closely and will share the facilities of a new building being constructed for them under the auspices of the Volkswageon Foundation, to be ready in 1968.

Dr. Cremer foresees two major problems for the future: first, the need for development of a national information system encompassing both subject- and mission-oriented centers, and secondly, the promotion of research work in documentation, with a concommittant need for better training of documentalists [53]. With respect to the first problem area, the establishment of "documentation rings" (e.g., for such subject areas as pedagogics, transportation, nuclear energy, agriculture, and nutrition) and central libraries (a central technical library at Hannover and central libraries for special fields, such as that for economics at Kiel). In West German industry, some interesting cooperative projects are being developed for the areas of chemistry, pharmaceuticals, and electrical engineering. In particular, cooperation between the major chemical firms has led to the formation of the Research Association for Chemical Documentation. It is hoped that cooperation will also be developed on an international basis, especially in terms of collaboration with the

Chemical Abstracts Service in the United States.

Training activities sponsored by the Institüt include special courses for post-graduate students (for example, on recent advances in reprographic techniques), and the training of medical assistants or technicians with respect to the transcription of handwritten records into machine-useful form. Although the Institüt für Dokumentationswesen itself is primarily a small "think" organization, its own research activities include application of computer techniques to real-time record processing in clinics and to related problems of data documentation. The Institüt is also providing support to university medical facilities for the planning of a system for the central evaluation of medical records.

Dr. Cremer reported that Dr. Pflug, the Librarian at Bochum, has been investigating possibilities for combining "peek-a-boo" coordinate indexing techniques and computer processing with respect to a multi-level selection and retrieval system. Dr. Cremer himself serves on the Library Committee of the German Research Association, which promotes a system comparable to the Farmington plan, a Union List of foreign periodicals, Union catalogs, and inter-library loan services.

Mr. Klaus Schneider is the director of the Zentralstelle für Maschinelle Dokumentation, which is also located in Frankfurt [54], [55]. As noted above, ZMD is a quite new organization, founded in 1964. The first programmers came on duty only in the last months of that year. In 1965, only a few computer projects were undertaken, but a major accomplishment was the mechanization of the preparation of the indexes for the German National Bibliography. Beginning in January, 1966, 26 numbers of Series A (lists of publications generally available at bookstores) and 26 of Series B (papers and other publications not as generally available) have been issued on a weekly basis. The West German national library, Deutsche Bibliothek, prepares the input citation material on 8-channel punched paper tape, using some special character-symbols to indicate significant words for keyword-index compilations.

This input material is presently processed through an IBM 1460 system to produce author indexes, publisher indexes, bibliographic citation data, keyword-in-context listings by previously marked input words, and the like (Fig. 14). (Schneider remarks that "pure" KWIC indexing presents difficulties for German language materials, first because of the language structure, and secondly because an adequate "stop"- or "purge"-list is extremely difficult to construct in view of the highly combinatorial word structures typically found in German.) During the first half-year of operation, some 14, 000 titles were processed, from punched paper tape input exceeding 84, 000 kilometers in total length. In addition, a quarterly listing is required for map-acquisitions, and cumulative indexes are required monthly, quarterly, semiannually, and by five-year cululative periods.

The differences between manual and automatic compilation are well reflected in the fact that there is now only a two-and-a-half month lag between the close of the semiannual cumulation period and the delivery of printed, bound copies as processed through the machine system, in contrast to the two-and-a-half years typically required for manual compilation in the previous system. Even this speed-up factor is seriously affected by the delays imposed by computer output to printing, and by binding requirements. Thus, output from the IBM 1460 is at the rate of 150 characters per second in the form of 6-channel punched paper tape that in turn feeds Linotype equipment (Linoquick) operating at a rate of only 3 characters per second.

The KWIC-type index produced (more properly, a "KWOC" or keyword-out-of-context index) consists, on a semi-annual basis, of bibliographic citations arranged by two types of keywords: the stichworts (derived automatically from the titles) and the schlagtwort, assumed to be the most important word(s) descriptive of the subject content, which are assigned by cooperating librarians for the documents in each month's cumulated receipts.

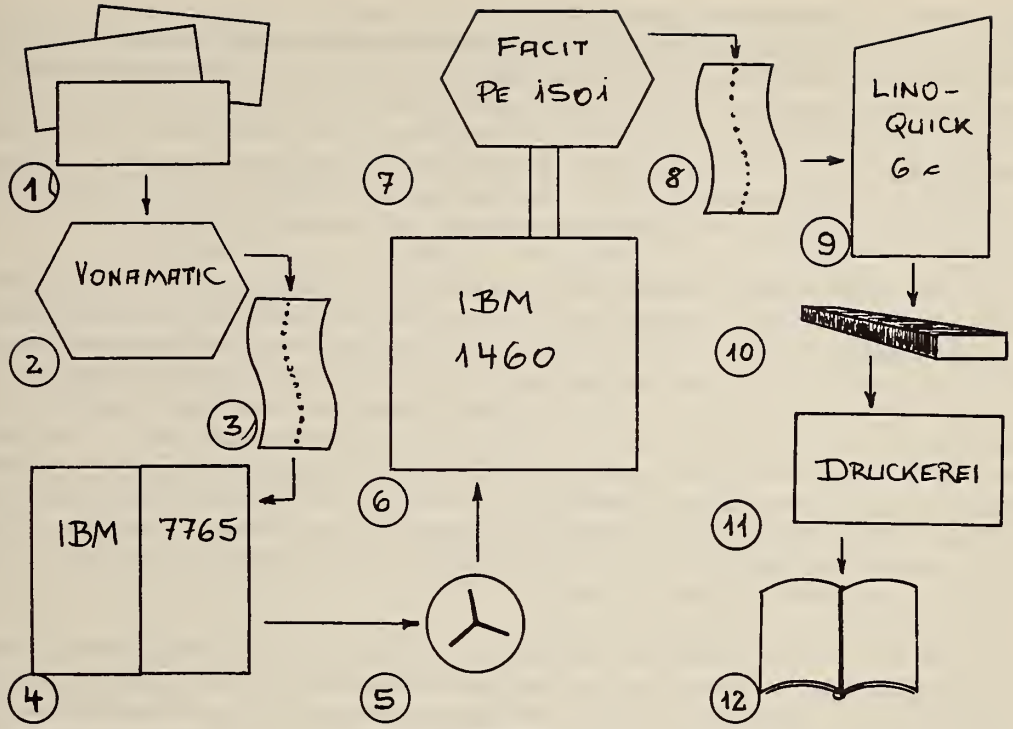


Abbildung 1
Technische Kette

1 *Titelaufnahme*
2 *Lochstreifenschreibmaschine*
3 *8-Kanal-Lochstreifen BCD-Code*
4 *Lochstreifen-Magnetband-Umwandler*
5 *Magnetband*
6 *Datenverarbeitungsanlage*
7 *Lochstreifen-Schnell-Locher*
8 *6-Kanal-Lochstreifen TTS-Code*
9 *Zeilensetzmaschine*
10 *Gußzeilen*
11 *Druck*
12 *Bibliographie*

Figure 14. Index Preparation System, Zentralstelle für Maschinelle Dokumentation

Random-access storage discs are used to store pertinent titles by schlagtwort assignments. (Some 14, 000 titles in the first half year of operation involved some 100, 000, 000 characters in machine-usable form). The output format presents the "keyword" left-most, with its right-context to the end of the line, and wrap-around to the succeeding line of print-out.

The ZMD is now working on an automated documentation system for the Center for Aeronautics and Space Flight Information, Munich, and on another documentation application for the Center for Mineralogy. A start has been made on mechanization of UDC indexes. Other tasks involve programs to convert data from one machine language to another. The special, continuing aim of ZMD is to provide technical assistance to other documentation centers in the mechanization of their operations. Because of a general lack of computer programmers familiar with nonnumeric data processing problems, special training courses are offered at the ZMD.

Also in the area of automatic documentation, the German Computing Center (Deutsches Rechenzentrum, Darmstadt) is involved in the planning and testing of systems ranging from simple word-indexes through acquisition announcements and special bibliographies to S. D. I. (Selective Dissemination of Information) projects in such subject fields as medicine, biology, agriculture, history, and sociology. Members of the staff serve as consultants in library automation for two or three German libraries and, in particular, they provided systems planning and programming assistance to the Bochum Library in the development of computer-compiled catalogs, automated circulation procedures, and the like.

Two special mechanized documentation projects are currently in progress at the Computing Center. The first of these is in the medical field and it involves the automatic processing of autopsy reports. These reports are provided with well-defined data sequences, but no organized format, on punched paper tape. The medical language used is a relatively standardized subset of natural language. The objective is to translate each report sentence into at least one code of a facetted classification scheme. A computer-based dictionary is used to define synonymous expressions and to provide some syntactic information as well. The facetted classification scheme used is flexible and can be extended to provide both new hierarchies and new categories. Some role indicators are also used. Updating can be carried out easily --- for the classification scheme, for the stored syntactic information, and for the indexing itself.

The second special documentation project at the Deutsches Rechenzentrum is in the field of the history of the arts and archaeology. In addition to the documentation of the pertinent literature, the clients want also to retrieve reproductions of the works or artifacts, so that it has been necessary to develop a language to describe potsherds, pictures, and the like. This descriptive problem has been divided into two phases: on the one hand, all the information that is geometric in nature or that can be described in terms of geometric properties; on the other hand, "everything else". For the latter, a special language, "Hisdoc", has been developed [56], providing a facetted description of the data of historical interest, where the facets are either factual categories or syntactic categories defining, for example, which referrents are the subject, verb, and object of a depicted action. Some 60 different categories together with some special connecting symbols are used in the language, which is claimed to be more sophisticated and powerful than that developed earlier by Gardin for archaelogical subjects.

With respect to geometric properties of items covered by the collection, a pencil-following device has been devised to track photographs of objects taken from three directions for computer processing in order to prepare 3-dimensional plottings and to determine coordinates of equidistant points along the curves. The projection curves are classified by topological scheme. They are size-normalized and then compared with the derived property measurements of other objects to determine classes of objects. Clustering techniques are used for these automatic classification purposes.

5.4 The Swedish Defense Institute

Laborator Folf Moore, of the Swedish Research Institute for National Defense, and more specifically head of the FOA (Försvarets Forskninganstalt) Index, is concerned with documentation research generally, with mechanized selection-retrieval operations more specifically, and with selective dissemination potentialities for relatively small and relatively homogenious clientele subsets in particular. He is actively cooperating with other Swedish organizations (e.g., use of the formal, natural-language-subset grammar developed by Ellegard at the University of Gothenburg) and will participate in the cooperative use of the IBM 360/75 time-sharing system which was scheduled to be installed in Stockholm in 1967.

Moore is concerned, first of all, with the modelling of documentation systems and particularly with the question of why systems that have worked well in the past are no longer effective today. He suggests, for example, that there are ten functions to be considered in models for all document systems. A particular topic of investigation relates to the question of optimum size of a clientele set or subset to be served by a particular system or service, and it is suggested that this optimum size is in the neighborhood of 30-50 clients for "custom" services. As a result, Moore envisaged several different levels of service. First, use of information available from international abstracting services, via either printed indexes or magnetic tape recordings. Next, there is a proposed intermediate level in which input from various sources would be combined into classes based upon various points-of-view for selective dissemination to variously defined, small, user-groups. Thirdly, a word-extraction process would be applied to establish criteria for selection-retrieval requests and to break a given mass of input items into overlapping, specialized indexes and current-awareness notifications.

In the present FOA system, manual subject content analysis results in the preparation of "retrieval sentences", following a standardized grammatical construction, to show the qualifying relationships between descriptors or keywords. For index production pruposes, each such sentence is permuted and listed under each of its keywords. For machine search purposes, however, keywords are coordinated without regard to relationships, as in straightforward coordinate indexing systems. However, on output, selected items are listed not only with identification-location data but also with their retrieval sentences so that the user may quickly spot the false or unwanted coordinations. It is of interest to note that the "man bites dog" cases are not screened out before delivery of results to the client because in some instances it has been of value to suggest, for example, "use of submarines to attack torpedoes" as well as vice versa. In addition, the user learns to provide better retrieval sentences for more effective searches than those that have been conducted on the basis of his original request.

FOA index listings are computer-produced by author, corporate author, report number, classified subject, and keywords in the retrieval sentences, which may include author name and year of publication as well as the content statement. Sorting is performed easily because input card identifiers use machine symbols in the desired machine-sort order.

Computer programs include "CORSAIR II", for keyword coordination search, a Retrieval Sentence program and a program under development for handling both types of search. These are written for the IBM 7090, but will be shifted to PL/1, COBOL, and/or TRAC for use with the 360 system in 1967.

Also under development are programs for "universal" input for conversion to standard machine code, standard tape format, and the like. This problem is of particular interest in Sweden, where some 30-40 different codes are used in various card and tape equipments. To date, several sections of code conversion programs (e.g., 8-channel paper tape to 1401 5-channel to a FORTRAN character set), two for standardized tape formatting, three or four processing programs, and several presentation routines are operational and they can be combined in over 20 different combinations.

5.5 Soviet and Other Examples

At VINITI (the All-Union Institute of Scientific and Technical Information), U.S.S.R., a demonstration selection and retrieval system has been developed and was discussed during the Symposium on Mechanized Abstracting and Indexing. The demonstration system involves approximately 25,000 documents in the field of electrotechnology and is designed for feasibility testing of a highly centralized service to the industry that would be capable

of handling 100,000-200,000 new items per year. To date, about 20,000 of these documents have been analyzed and indexed by machine, using thesaurus look-up to identify related terms, synonyms, and the like, and to provide word-by-word translation of document input to the numerically encoded documentary language which is that of a coordinate descriptor search system, involving 3,500 descriptors at the present time. Up to now, only short documents (i.e., abstracts) have been punched and processed and it was noted that present machine capacities limit item processing to 4,000 characters on average.

Difficulties of translation to the documentary language code so far encountered have included: (1) the question of inflections of Russian words with problems of both machine capacity and processing time, (2) the fact that words and expressions connected as to concept may be separated in the actual text so that it is necessary to detect the possibility of such connections --- e.g., whether "device" is a single word or is part of a phrase such that it is to be translated as "voltmeter", and (3) the problems of homography. Some 26 homographs have been identified as particularly troublesome with respect to system noise for the literature of electrotechnology and three of these have been analyzed in terms of lexical context only (i.e., no syntactic analysis has been used).

A machine-based, 4,000-item, Russian word stem dictionary is used. For each text word, the appropriate stem is selected. Words not found in the dictionary are printed out, so that misspellings and possible new words can be detected. Print-outs may also be made of possible beginnings of complex expressions (e.g., where "installation" or "resistance" may be the beginning of a phrase rather than a single word). If a word found in the dictionary is not marked as being a possible "complex" or a homograph, it is translated directly into the 4-digit concept code.

If the word is a homograph, a sub-routine examines established lists of co-occurrences of specific words with the various meanings of the homographic word and looks for these words differentially in the contexts of three words to the left and five words to the right. For the three homographs so far handled by the present program, 90 percent of the homograph separations are correct, 4 percent are incorrect and 6 percent involve cases that cannot be resolved. The selected codes for each item are arranged sequentially and duplicates are eliminated.

The documentary language involves the use of the concept-codes or descriptors as manually established, where the code structure directly incorporates certain semantic relations, including some hierarchical relations, between descriptors. In addition, manually determined links or connections between descriptors have been set up for retrieval purposes, tested by experimental searches for the total list of descriptors conducted in parallel by machine and by subject specialists, and refined for machine use on the basis of analyses of noise, omissions, and desirable improvements.

The retrieval logic depends upon these connections in the sense that if a query term does not find a direct match then search is made for match on related terms. If any descriptor in the question finds no direct or related term match with the stored item, the item is not selected.

Output of references for selected items is in two listings --- those which the machine considers positive responses and those which because of tracing of connections to presumably related descriptors are only probable. The machine also prints out listings of the connections that were used so that statistical analyses may be made as to the extent to which proper use is being made of the connections. Very preliminary results indicate that for 100 relevant documents in the collection, about 55 are selected by direct match and 25 by virtue of the pre-established connections. It is hoped that the latter procedure can be improved to 30 or 35 percent, to give up to 90 percent recall.

Early manual simulations gave encouraging results for a 2,000 document subset. More recently, 30 questions have been run by machine against 10,000 items. The jump from 2,000 to 10,000 items did not significantly alter the performance. Noise amounts to about 40-50 percent of the output.

Also at the Moscow Symposium, current work in mechanized documentation at the International Atomic Energy Agency, and Vienna, projects in Czechoslovakia, were reported. K.V. Ivanov and G. Del Bigio described the documentation activities of this international organization, which has 96 cooperating national members. An information storage, selection and retrieval system for atomic physics has been under development since late 1964. The present store contains approximately 12,000 documents. A 1401 computer was acquired last year. Every two weeks a list of references for new acquisitions is prepared and annual indexes for 1964 and 1965 have been issued. A master program for KWIC indexing provides variable formats and pagination. Special indexes, lists, and bibliographies are also prepared. Twenty members of the Agency are involved in planning for a center for which a larger computer will be acquired next year.

Mrs. A. Veisova reported on documentation activities in Czechoslovakia. Since 1963, there have been some 20 projects at the Institute for Science Technical and Economical Information, 14 involving punched card equipment and 6 running on computer. She gave as a first example a collection in the field of geology with about 4,000 new items per year. She also described experiments in KWIC indexing for input materials in German, English, Russian and Czech languages, but reported that at present only English is used because of the linguistic problems encountered with the other languages. Multiple copies of this English KWIC are prepared by xerographic techniques. Brief mention was also made of a Union Catalog of translations, a program clearinghouse, and projects involving network analysis of the national information system.

In terms of KWIC-type indexing, Schneider of the Zentralstelle für Maschinelle Dokumentation (ZMD), Federal Republic of Germany, also emphasized the difficulties of developing an effective stop-list for German language materials. He had previously noted the difficulties of designing adequate "stop-lists" for the processing of German language titles because of the problems in the multiple compounding of German words.

It is also to be noted that at Ceccato's Centro di Cibernetica e di Attività Linguistiche, University of Milan, a project in documentation of legal literature involves a combination of manual and machine-aided classification as well as studies of the comparative effectiveness of decimal classification and keyword indexing.

Another project at KVAL (Research Group for Quantitative Linguistics), in Sweden, is concerned with investigations from a mathematical point of view of efficient ways in which information about bibliographic citations may be used for literature search, with emphasis upon a proper balance between "citedness" and "citingness" relationships among the documents in a library. The objective of the investigators, B. Brodda and H. Karlgren, is to find computer algorithms that will replicate effective citation tracing as done by the human searcher [57].

Karlgren is no longer directly concerned with automatic typesetting procedures although, like Allén at Gotherburg, he feels that analysis of intra-word grammars (e.g., those of phonemic combination) are important to the resolution of problems of hyphenation, especially to some difficult problems encountered in the Swedish language. However, the KVAL organization continues to provide consultant services to AUTOCODE AB, a Stockholm firm engaged in commercial development of automatic typesetting techniques, with a computer typesetting program in ALGOL available.

6. LINGUISTIC DATA PROCESSING

One of the earliest of the possibilities for nonnumeric data processing to be explored anywhere was that of using machines in the area of computational linguistics, specifically, the making of concordances and the carrying out of word frequency counts. Similarly, from the early days of computer technology, there has been continuing interest in the possible use of machines for automatic data processing of natural language texts.

6.1 Computational Linguistics

At the University of Gothenburg, Sweden, work in computational linguistics has been pursued by A. Ellegård, S. Allén, and others for some years [55]. For example, in 1961-1962 a concordance to a 17th century Swedish text of 44,000 words was produced, using punched paper tape input and an ALWAC III-E drum computer. By 1964, the program was rewritten for the SAAB D 21 computer in the DAC programming system by Dr. Sixten Abrahamsson, to serve as a general-purpose concordance preparation and word frequency listing procedure.

A current project is concerned with word frequency studies of modern Swedish newspaper text, under a grant from the Bank of Sweden Jubilee Foundation to establish the Research Group for Modern Swedish. The corpus will consist of approximately 1,000,000 running words in 1,500 articles in the form of 6-channel teletype tape received from the typesetting operations of five cooperating newspapers during 1965. After modification of the SAAB D21 tape reader to accept this input, running numbers are assigned to each article, the teletype code is translated into machine code, hyphenated words are recombined, line-ends are marked, and this converted text is stored on magnetic tape.

The frequency lists available after processing include sums of running words; sums of different graphic words; sums of alphabetic, numeric, and hybrid (e.g., "1900-talet") character combinations and junctures. Programs prepared by Abrahamsson and Hagstrom are available both in DAC and in Algol-Genius. Special studies to be undertaken will include deriving statistics for average word and sentence lengths, determining the frequency distributions of various parts of speech, and investigating possibilities for computer-based syntactic analysis.

In discussions with Dr. Ellegard and Dr. Allén, it was learned that current work is devoted to study of frequency lists of morphemes and to analyses of how morphemes combine into words. It is hoped to go on from frequency studies to syntactic analyses, but it is suggested that analysis of deep, rather than merely surface, structures is as yet a long way off.

At the Institut für Phonetik und Kommunikationsforschung, University of Bonn, Dr. D. Krallman's interests in automatic language processing and information selection and retrieval are presently directed principally to investigations of general methods for language processing [59], [60], [61]. In collaboration with Dr. Martin, of the Philosophy Department of the University, and with other Institutes, such as that for philology, statistical analyses of text are carried out, e.g., to count features and compute various measures of interest in problems of stylistic considerations. A major corpus consists of the first nine volumes of the collected works of Kant, with 1,400,000 text words, 57,000 different word forms, and 20,000 different lexical units. (The derivation of lexical units from word forms is performed manually.) A system of multi-purpose linguistic analysis programs has been developed, and texts in Spanish, French, Middle High German, and from East and West German newspapers are being processed for statistical and semantic comparisons. (Two keypunch operators are kept busy eight hours a day preparing additional text.)

Also at this Institute, Dr. H. Schnelle's interests in general linguistic methodology relate first to research into the grammatical description of German word formation, e.g., to find rules describing the restrictions on combination and to investigate the role of suffixes and prefixes in transforming a word from one category to another. Word stems are classified into the traditional word classes; affixes are classified with respect to their transformational effects. It is then possible to derive a generative procedure for legitimate word combinations. Schnelle has provided a description in the form of Bar-Hillel's categorial grammars, and a corresponding program is being written (in the COMIT language for the 7090) that will include morphological and inflectional processes.

In addition, Schnelle's work in general methodological research includes: (1) study of the relationships between formal grammars and the structure of automata; (2) investigation of "Cycle-Coordinated Automata" [62] into which CF (context-free) grammars may be translated by deriving a graph, corresponding roughly to Gorn's flowgraphs, from the structure of the grammar's rule system; (3) the study of propositions of the form: simple predicates, strings, predicates and strings, to build up metalinguistic expressions; (4) the introduction of lexical propositions [e.g., "'the' is an article", "'man' is a noun", "if a noun phrase 'the man' occurs, '_____' is a possible predicate"], and (5) use of combinatorial logic as a starting point for better formalizations of propositional inferences. In much of this work, there is an emphasis on context sensitivity. Other dissertation and thesis work is being conducted at the Institute and there is also a project to translate from orthographic language into phonemes by computer, with a first program written in FAP and another being written in COMIT.

At VINITI, Moscow, a system of programs has been developed for linguistic data processing. These provide for translations from text to a "vocabulary of grammatically connected words", for a KWOC [keyword-out-of-context] index, for "linguistic commentary adaptive to translate words by frequency data", and for work on historical texts. In the latter area, work has been done on deciphering a language from about 3,000 B.C. as to whether it is of the Davidian type. However, the same technique applied to Easter Island text was not successful because of the frequency of abbreviations that were encountered. In the VINITI linguistic research program, investigations are also being made with respect to probabilities of words and grammar classes in text. An approach similar to that of Mandelbrot is being pursued, but with some differences. The present model is such that Zipf's Law can be deduced from it.

6.2 Automatic Syntactic Analysis and Other Applications

Professor Funther Wenck, Seminar für Sprache and Kultor Japans, University of Hamburg, West Germany, has been concerned with possibilities for automatic translation of Japanese language texts. Under a EURATOM research contract he has studied translation of Japanese scientific texts into English, including problems of handling Japanese ideographs with a view toward automatic processing of graphic data. However, be reports that the main emphasis at the moment is on the completeness of the linguistic criteria to be applied in syntactic analysis rather than on equipment considerations.

The Language Research Section, IDAMI (Instituto Documentazione della Associazione Meccanica Italiana), headed by Dr. Ernst von Glasersfeld, was visited shortly before the relocation of this organization from Milan to the United States, with P.E.C. Research Associates, Inc., Athens, Georgia. Their recent work [63], [64], has been primarily directed to the development of a syntactic analysis program for 10-word English sentences, presently programmed in machine language for a GE 425 system with 32K memory, and to be re-programmed for an IBM 360 system. One-sentence strings only are now processed, with necessary dictionary and syntactic information added from a punched card file since

limited peripheral facilities prohibit computer storage at this time. To date, 250 syntactic functions have been identified, and the procedure handles about 90-95 percent of English syntax constructions. The iterative analyses take at most two seconds, the majority being carried out in less than a second. Other topics investigated have included studies of the functions of articles and prepositions in English [65], [66], [67] and the development of "Multistore" and other procedures for correlational analysis [68], [69].

Dr. Hans Karlgren is affiliated with the Swedish Board for Computing Machinery and KVAL, the Research Group for Quantitative Linguistics, in Stockholm. In addition, he is the editor of SMIL (Statistical Methods in Linguistics). At KVAL, he and Dr. Benny Brodda and a small staff are working on various projects involving linguistics, mathematics, and computer sciences, supported both by the Swedish Government and by private funds. The present work in linguistics is concerned more with recognition than with generative grammars, starting with sentences as given in a text to find their syntactic structures and exploring a categorial grammar approach after Bar-Hillel. Prior work in compiler construction is being applied to string analysis either by computer or my manually-simulated, but potentially useful methods.

These investigators have been particularly interested in problems of man-machine interaction, especially for information selection, storage, and retrieval systems, stressing the need for better constructed, more man-like, computer languages for this purpose. Linguistic studies related to this concern include studies of nominal phrases, especially nested phrases, in order to determine which ones can be analyzed formally and which ones contain hidden ambiguities. One objective is to choose a subset of language for the man-machine dialog so as to avoid unanalyzable or structurally ambiguous constructions. In addition, they have applied mathematical analysis procedures to problems of synonym detection and the construction of thesaurus classes, with a computer program starting with one word to build such classes. Highest-class-score selection methods may be applied in further text processing.

Linguistic data processing work at the Deutsches Rechenzentrum (German Computing Center), Darmstadt, began with straight-forward lexicographic analyses (concordance-compilations, word frequency tabulations). In particular, the "Index" program, which has been operational for five or six years, provides 15 different index-compilation options [70], including capabilities for sorting out incidences of proper names and a morphological analysis routine in which extractions are made on the basis of specifications of any desired sequences of characters (Fig. 15).

Continuing linguistic data processing research includes some work in automatic syntactic analysis, involving a mixture of elements from various approaches suggested by others. Members of the Computing Center staff have conducted analyses of English sentences in articles from the London Times, and it is reported that the system works well, although it is slow and requires a fantastic amount of previously stored information. They have also tested various possible transformational rules by using them as input to a sentence-generator program and they have done some phonetic work as well, in which, instead of sentences, phonemic sequences (in English) are generated that are syntactically correct. The latter project requires transcription of the international phonetic alphabet into the machine-usable 48-character set.

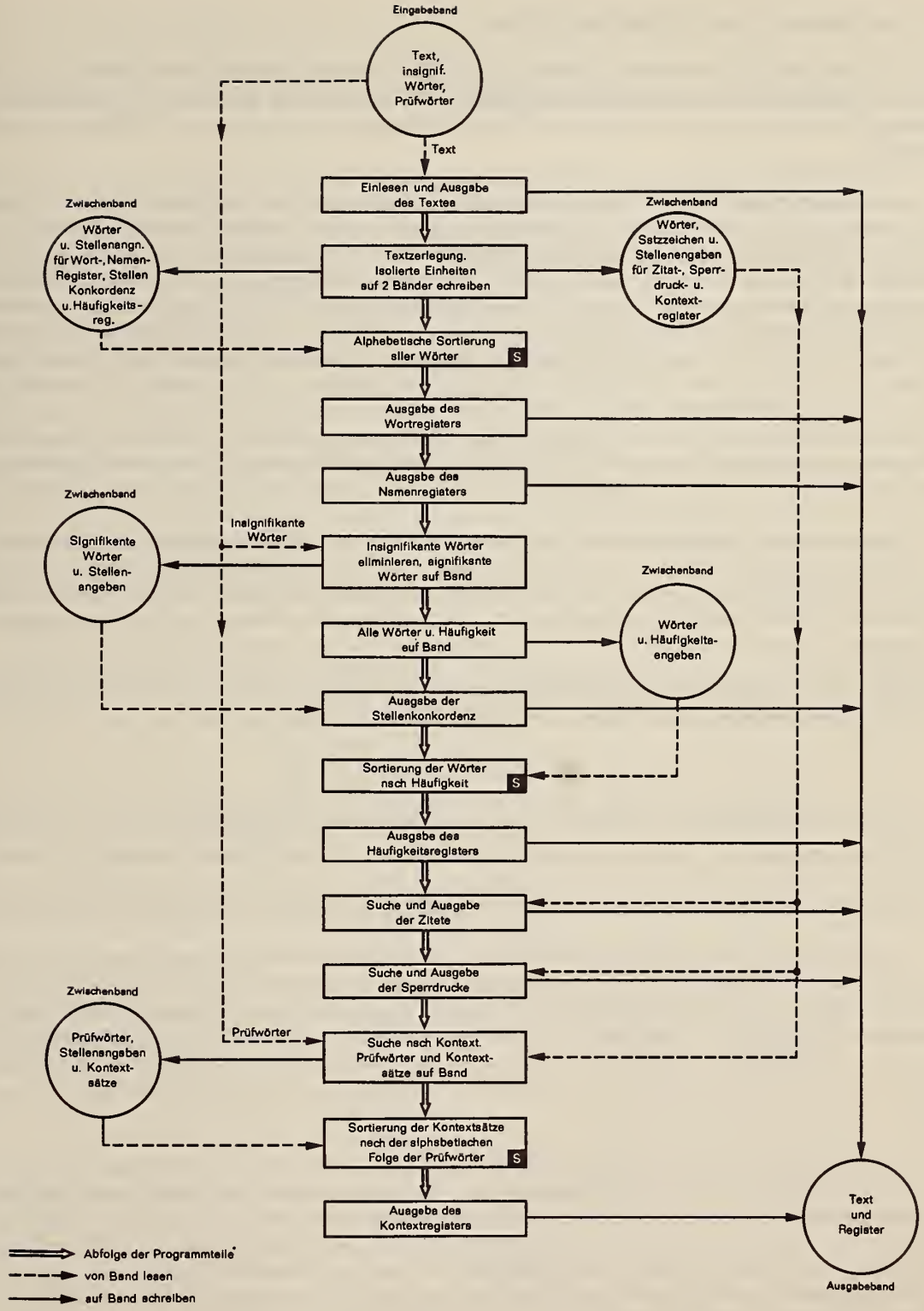


Figure 15. Block Diagram of the "Index" Program, Deutsches Rechenzentrum

6.3 Mechanized Abstracting and Indexing

Experiments in automatic abstracting and/or indexing were noted, first, at the National Physical Laboratory, Teddington, England, and secondly at the Moscow Symposium. In addition, results of work by Wagner of the Technische Hochschule, Karlsruhe were made available.

Automatic indexing and documentation research continues at NPL. Peter Vaswani and Roger Meetham discussed their current research. As a result of earlier activities [71], [72], word-pair association data had resulted in the development of a vocabulary of 1, 000 possibly significant "clue-words" in the form of word-stems. An association measure based upon conditional probabilities has then been applied to the analysis of 12, 000 abstracts in 5 subject categories (automation, computers, circuits, general physics, and geophysics.) A sample of 1, 500 abstracts was used to obtain a 1, 000-word-stem "reference dictionary" (where, on average, match may be made for four different word endings.) When keypunched text for the 12, 000 abstracts was processed through the ACE computer, the keywords were selected by matching against this reference dictionary, and both word-frequency and co-occurrence data were obtained. For any two words occurring in the same abstract, data with respect to the number of abstracts in which they co-occurred was specifically obtained.

Experimental computations have shown that for simple conditional probabilities [p(a/b, p(b/a)], a symmetric function threshold did not yield some important pairs, especially if one word "a" has occurred in many different contexts, and word "b" has occurred only in a highly specialized one. Nevertheless, because of drum storage-access requirements, a symmetric access was desired. Thus, correction factors may be applied. If x (p_1, p_2) is desired, then it may be obtained by taking into account the extent to which the individual conditional probability exceeds the unconditional probability:

$$p(a/b) - p(a) = p_1.$$

Similarly, for p_2 then, the greater of these two corrected probabilities can be used as the association measure. Actual association values for a 1000 x 1000 association-value matrix are converted to a binary matrix by appropriate threshold settings. Similarities of common "members" for pairs of rows, are then calculated to produce similarity coefficients.

Graphs may be obtained by representing clue-words as nodes with the connections between them weighted to correspond to statistical data with respect to prior associations. Present research is directed to the devising of algorithms for purposes of identification and isolation of subsets of nodes that are strongly interconnected. Several different approaches are being investigated. Exclusive groupings are not desirable, because of the different contexts in which the same word will be used. How large the sets should be has not yet been determined.

Present plans for implementation, testing, and evaluation call for the indexing of the source items exclusively. It is thought that, because of the large size of the sample (with 12, 000 items), the source data bias will not strongly influence the results. (However, some non-source items may also be processed later, to check this hypothesis.) Twenty subject specialists have been cooperating in the project and have together produced a set of 100 search requests, based on original source documents. Retrieval selection is by conventional descriptor matching for n, n-1, etc., descriptors, resulting in a rank-ordered output list. Output items exclusive of the item upon which the search request was based are referred to the original "requestor" for his subjective estimates of relevance. His relevance assessments are fed back into the system so that on later runs he will not be sent items which he has already rated.

In the automatic indexing work at Karlsruhe, Dipl. -Ing. S. W. Wagner reports that his approach is somewhat similar to that of Edmundson and Wyllys with respect to the automatic extraction of index words from full text but that, instead of relative word frequencies as the basis for selection, he uses rank criteria, with a specified interpretation of frequency-rank-distribution [73].

The Unesco-VINITI Symposium in Moscow was expressly devoted to the problems of automatic indexing, automatic abstracting, and related areas of linguistic data processing research. The three basic working papers commissioned for the Symposium were as follows: "Progress and Prospects in Mechanized Indexing", M. E. Stevens; "The Problem of 'Auto-Abstracting'", M. Coyaud, and "Studies on Automatic Indexing and Abstracting in the USSR", A. I. Mikhailov.* In general, while there was some optimism with respect to automatic indexing (of both the derivative and the classificatory types), there was considerable pessimism with respect to the practicality of auto-abstracting techniques at the present time.

Following M. Coyaud's presentation, the specific question was asked as to whether auto-extracts are being used in the USSR. Dr. G. E. Vleduc reported that there have been experiments only and commented on the prematurity of attempting to evaluate results in the light of the lack of standards for evaluating manually prepared abstracts. Other comments by Soviet participants included reference to the needs for fundamental research not only with respect to the bases for "semantic descriptions" but also as to what types of abstracts now exist and how they are used. Reference was also made to an experiment applying automatic syntactic analyses to Russian and Ukrainian texts for auto-abstracting purposes.

Dr. Mikhailov's presentation stressed first the desirability of greater international cooperation in the field. He then emphasized the two aspects of the problems, the theoretical or scientific on the one hand and the practical on the other. With respect to the theoretical side, he stressed that the realization of automatic abstracting and automatic indexing in the strict sense (i.e., the translation of text to a formalized language, the identification of the main content, the expression of the subjects in formal language and the translation back into natural language) would be possible only after deep research in linguistics, psychology, mathematical logic, and semantics.

However, the solutions to these problems are of general scientific interest and will provide a key to the mysteries of human thinking. Although necessarily long-range, this research should be both widened and deepened. In Mikhailov's view, solution of the problems of machine translation is necessary to the development of successful automatic abstracting and indexing. The automatic reading of texts is also necessary, since without such techniques it will be impossible to achieve practical solutions. Indeed, such techniques are necessary for research in the field.

The problems of satisfying the information needs of scientists of today are very complex, so that one cannot expect to arrive at practical solutions in the near future. On the other hand, one cannot entirely wait for the development of pragmatic methods until the theoretical problems have been solved. In order to reach solutions, many methods must be explored; however, the elaboration and use of methods for practical solutions should not hold up basic research.

The UNESCO Summary report on the Symposium [74] emphasized the following points:

(1) "The conclusion was that, while there is no reason to assume that automatic abstracting is impossible, it must be regarded as a goal of long-term research efforts. In the ensuing discussion, the overall view expressed towards auto-extracting

* The Proceedings of the Symposium are in process of preparation but are not yet available.

was somewhat pessimistic. The participants were hopeful that auto-abstracting would one day become feasible, but stressed that much basic research remains to be done. Similarities were pointed out between the problems involved in auto-abstracting and those arising in connection with machine translation of natural languages. ...

(2) "One of the most important problems, and a serious inhibitor of the research efforts in natural language processing, is the lack of a suitable body of machine-readable texts and lack of adequate character-recognition devices capable of reading several kinds of type fonts. ...

(3) "Because of existing needs for fundamental research on linguistics, Unesco should establish a working group on computational linguistics ... to dtudy and prepare a report on those aspects of linguistic analysis, particularly transformational analysis and syntactic and semantic models of languages, which are relevant to automatic indexing and/or abstracting. ...

(4) "With particular emphasis on goals which can be achieved in a reasonably short period of time, Unesco should consider establishing a working group which will ... suggest to a number of professional societies in various countries and to journal editors that they print or otherwise provide a list of all alphanumeric and special characters that have been used in type-setting the journal. Such a list would be a valuable aid when preparing to encode these texts by keypunch or by character recognition devices."

7. COMPUTER CENTERS, COMPUTING AND PROGRAMMING THEORY, AND MISCELLANEOUS PROJECTS

Computer centers were visited in the Netherlands and West Germany, especially the Deutsches Rechenzentrum at Darmstadt. In addition, a number of computing or programming theory and miscellaneous other projects were noted in the course of the survey.

7.1 Computer Centers

At the Netherlands Mathematics Center, in Amsterdam, P.G.G. Van de Laarschof is primarily interested in numerical analysis and computer programming, especially ALGOL compilers and translators, and in the use of the LISP language. An automatic typesetting program is available in ALGOL-60. The Center has a new computer, X-8, especially designed to operate with ALGOL programs. It has 12K memory, 2 1/2 μs access time, and drum auxiliary memory of a half million words.

Dr. F. Schulte-Tigges is the director of the nonnumeric data processing program at the Deutsches Rechenzentrum (DRZ, or German Computing Center.) This institute was jointly founded by the government of the Federal Republic, by the community of West German States, and by the West German equivalent of the National Science Foundation, in 1960. It serves as a computing center for the German universities (Fig. 16) [75], especially for large-volume data-processing problems where relatively small computers are not adequate to the task. As such, the Center is therefore committed to large-scale, high-speed equipment. Starting with an IBM 704, they have progressed through 7090 and 7094 installations, with a 360/50 system expected in November 1966, and with plans for future expansion to a larger 360/75 system or possibly to the Telefunken TR-40 computer system.

Figure 16. Network of Cooperating Universities, Deutsches Rechenzentrum

As a non-profit organization, the Center works free of charge, billing only a low fee ($60.00 per hour) for the actual running-time use of the computer installation itself. Some 80 different nonnumeric data processing projects are typically current at any one time, so that, with a staff of about 25 technical people, there are definite problems of scheduling and queuing. Professional training is also provided, especially with respect to nonnumeric applications. Programming courses are provided for about 100 people a year, drawn from graduate students and university faculties. Special seminars are also provided. Training courses for junior programmers, limited to young female high school graduates, and of two and a half years duration, involves four semesters of mathematics courses at the Technische Hochschule as well as practical in-house programming experience. In addition, professional symposia on various research and development topics are sponsored, such as one on the formal structure of speech[76].

Work on programming systems at DRZ started in 1961 with the logical parts of FORTRAN and then added packaged, hand-coded sub-routines and macros adapted to the Computing Center's specific requirements. (Typically, programs with which the Center is involved run to 10-40,000 instructions with large data volumes to be processed.) By 1963, serious troubles were being encountered with respect both to storage requirements and to total production time. A chain mechanism was then worked into the modified FORTRAN system, providing for selective call-up of any specified link in the chain. A paper by Reul and Miedel for the Ninth European SHARE Conference describes the use of FORTRAN II and FAP at the Rechenzentrum together with development of a string manipulation package and a generalized sorting system allowing for the sorting of alphanumeric strings in accordance with arbitrary alphabets [77]. Current developments continue to be aimed at the elimination of many of the existing macros by a classification of basic operators and the writing of a new, pre-processor, interpretative language.

Documentation projects at DRZ have been noted previously. Other Rechenzentrum projects in progress include a pattern recognition program for a 12 x 80 binary matrix, designed to find pre-specified patterns and to count occurrences of recurring patterns, and a program for input and correction of data or text from punched paper tape involving some editing and accurate line-number identification of areas of editorial changes. (It is to be noted, with respect to editing, that in such projects as the one for specialized archaeological documentation mentioned earlier, considerable man-machine interaction is involved in an intermediate editing process. That is, machine results are displayed for subject-specialist review, and he indicates which of the machine-generated results are correct, adding any peculiar properties of interest as derived from his own analyses.)

Dr. H. Zemanek, of the IBM-Austria computing center in Vienna, reports that his group is no longer concerned with research in self-organizing system and learning models, but instead has been concentrating for the past four years on formal languages and, in particular, on a formalization of PL/I.

7.2 Computing and Programming Theory

A number of nonnumeric data processing projects at the Technische Hochschule, Karlsruhe, have been mentioned previously in this report. Of special interest in connection with computing and programming theory, however, is the work of Dr. W.J. Görke, who is interested in adaptive systems and the redundancy-reliability characteristics of the circuits of such systems in terms of broad applicability to information processing and communication systems application. A project sponsored by the West German Ministry of Defense involves investigations of self-correcting translator circuits with matrix structure for error detecting and error correcting codes, and in particular, self-correcting circuits for Hamming codes with minimum distances of 2, 3, and 4 (Fig. 17) [78]. Görke is also interested in the hardware realization of failure detecting and self-repairing circuitry and has designed two very effective demonstration models where component failures and short-circuits can be simulated, but correct results nevertheless achieved. For example, a decoding circuit for Hamming distance 3 codes provides for the decoding of a one-error correcting code in the presence of a limited number of circuit component failures.

At the Centro Studi Calcolatrici Elettroniche, of the University of Pisa, a group headed by Dr. A. Carraciolo is pursuing a number of computing and programming theory interests in three principal areas: (1) Symbol manipulation languages, (2) simulation languages, and (3) the development of appropriate languages for use in automatic machine tool control. More specific recent interests include:

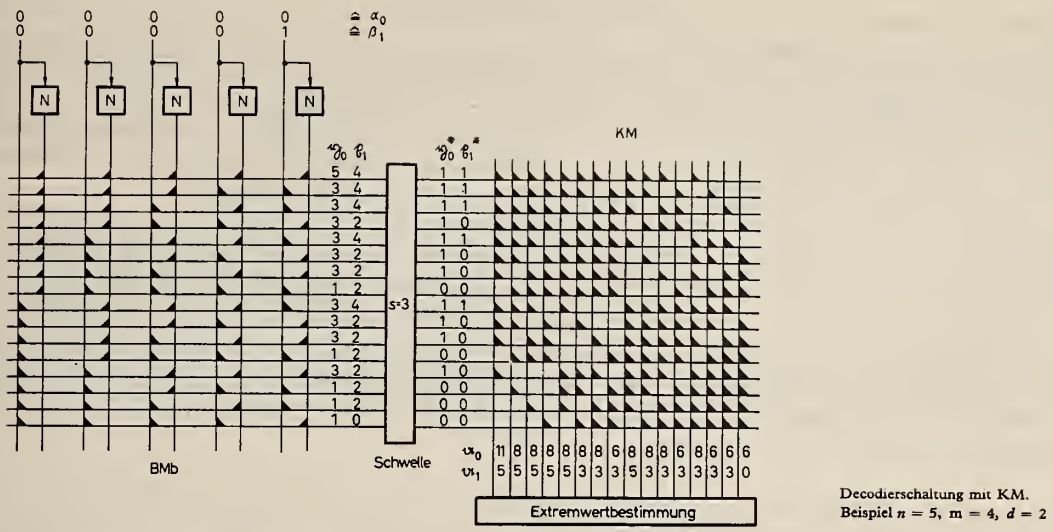


Figure 17. Self-Correcting Circuits for Hamming Codes

(1) Questions of linguistic problems in programming theory, including questions of the concept of a job and of the management of sequential and of parallel concurrent interacting jobs, [79] in particular:

"The theory of asynchronous processes is still not very advanced, but the same is perhaps true also for synchronous processes. The point in question is the following: Whereas in purely sequential jobs there can arise no contradictory requirements since each step is, by definition, uniquely determined; the situation is quite different for parallel jobs, where circumstances may arise in which a given facility is requested to perform at the same time two or more incompatible actions... Problems of this sort are of particular concern for so-called simulation languages and, therefore, represent an open question of the highest importance."

(2) The use of a generalized Markov algorithm in general-purpose programming languages, e.g., for more powerful pattern recognition systems, multistring manipulation, and self-adaptive capabilities [80].

(3) The questions of "theoretical pragmatics" --- that is, the relationships between a language, artificial or otherwise, and its users [81].

(4) The theoretical approach to the definition and formalization of special-purpose languages, such as those involved in machine tool control or in simulation experiments [82], [83], [84].

Among current developments at Pisa, it is noteworthy that a syntactic analyzer program has been developed (L. Spanedda is working on this) that, although somewhat similar to Chomsky's transformational grammars, deals with ordered sets of transformational rules. The questions of ordering (or, rather, of respect to invariance in terms of various possible orderings), is crucial to the adequate simulation of parallel, concurrent systems. The Pisa SL/1 language is based in part on SOL (Simulation-Oriented Language) and in part on SIMULA (the ALGOL extension developed by O. J. Dahl, of the Norwegian Computing Center, Oslo.) A second version, SL/2, now under development, will provide self-adapting features to optimize the system. Caracciolo emphasizes that, for any set of

deterministic processes which are to be applied simultaneously, but where problems of incompatibility may arise, the problems can be reduced to a set of probabilistic processes. Otherwise, if one sequentializes parallel, concurrent processes actually dependent upon the order of sequentialization, then hidden problems of incompatibility may vitiate the obtained results. He is also interested in the development of grammars for pictorial data description in which a basic superposition operator is both associative and commutative. This approach is suggestive for a number of pattern recognition problem areas.

The syntactic analyzer program, on which Spanedda is working, is used both for analysis of strings written in the "PANON" Language [85] and for experiments with Italian natural language grammar. The program is operable in two modes, either to provide a single decomposition or to trace all possible syntactic structures. Pilot trials are being constructed with the collaboration of a law school graduate student on a program designed for optimization of grammars, using Italian text, to eliminate redundancies of structure and develop a non-ambiguous grammar.

Prof. Onesto, also at the University of Pisa, is interested in areas of biological cybernetics and intelligence, problems of symbol representation, and the use of programming languages as clues to understanding more about phenomena of learning and memory. To date, however, he has not reported formal results in these areas.

7.3 Miscellaneous Other Projects

At the University of Naples, Italy, two groups are engaged in research pertinent to cybernetics and self-organizing systems. One, headed by Prof. Valentine Braitenberg, is concerned with the neurophysiology of systems, such as vision in flies, and with investigations of the cerebellum as a timing organism. The other group, headed by Prof. E.R. Caianiello, is primarily concerned with theoretical investigations and models [86], [87], [88], including: (1) studies of neuronal decision equations; (2) development of a theoretical model where decision elements are connected to others as nodal elements in a network and where the connections have both learning and forgetting dynanisms, with network responses at given time intervals recorded on punched cards; and (3) theoretical investigations of learning processes where the learning and forgetting mechanisms occur very slowly as changes of coupling coefficients.

With respect to problems of language, Caianiello stresses the redundancy and error-correcting facilities readily available in natural languages and he is exploring ways in which coding may be accomplished from characters to syllables (using as concrete examples 450 syllables in the Italian language) and from syllables to groups of syllables.

8. CONCLUSION

The survey on which the present report is based provided an unusual opportunity to visit over 80 nonnumeric data processing projects in Europe and to discuss problems of mutual interest with a number of research investigators in various countries. An important overall impression is that of both the high intelligence and the enthusiasm of the project leaders and staff personnel who were visited. On the other hand, limitations of resources (funding, time, adequately advanced and versatile machine and laboratory facilities, and availability of properly qualified manpower) appear to impose certain constraints on both the scope and the speed of European nonnumeric data processing accomplishments.

ACKNOWLEDGMENTS

The information upon which this survey report is based could not have been obtained without the active interest and support of Messrs. R. Wilcox and D. Pollock of the Information Systems Branch, Office of Naval Research, and of Mr. T. W. Marton, Department of Advancement of Science, Unesco.

In addition, the cordial hospitality and friendly cooperation of a number of individuals, most of whose names are mentioned in the text or references of this report, must be gratefully acknowledged. The clerical, typing, editorial, and bibliographic assistance of Mrs. Betty Anderson is also gratefully appreciated.

A somewhat condensed version of Sections 2, 3, and 4 of this report is in process of publication in Pattern Recognition, the Journal of the Pattern Recognition Society.

REFERENCES

[1] The British Computer Society, Character Recognition, London, 1967, 195 p.

[2] Coombs, A.W.M., Character Recognition (and Its Application to Postal Mechanization), preprint of a talk given at various centers of the Institution of Post Office Electrical Engineers, 1964-1965, 4 p.

[3] Clowes, M.B., The Use of Multiple Auto-Correlation in Character Recognition, in Optical Character Recognition, Ed. G.L. Fischer, Jr., et al, pp. 305-318 (Spartan Books, Washington, D.C., 1962), (See [10].)

[4] Clowes, M.B. and J.R. Parks, A New Technique in Automatic Character Recognition, The Computer J. 4, 121-128 (1961).

[5] Clayden, D.O., M.B. Clowes and J.R. Parks, Letter Recognition and the Segmentation of Running Text, Inf. & Cont. 9, 246-264 (1966).

[6] Klönne, K.-H., Automatic Check Handling in Germany, Elect. Commun. 40, 328-337 (1965).

[7] Braunbeck, J., Vergleichsmaske für Maschinelle Optische Zeichenerkennung, Deutsches Patentamt Auslegeschridt 1 207 685, Dec. 23, 1965.

[8] Dietrich, W., Optical Character Readers for Automatic Document Handling in Banking Applications, Elect. Commun. 40, 312-327 (1965).

[9] Gattner, G. and R. Jurk, Application and Technique of Automatic Character Recognition, Siemens Review XXX, 390-393 (1963).

[10] Gattner, G., Automatische Zeichenerkennung für Postleitzahlen und Gesprächszähler fotografien, Der Ingenieur der Deutschen Bundespost 3, 103-109 (1965).

[11] Institut für Nachtrichtenverarbeitung und Nachtrichtenübertragung der Technischen Hochschule Karlsruhe, Forschungsbericht über Automatische Zeichenerkennung Selbstrkorrigierende Schaltungon und Adaptive Systeme, Karlsruhe, Germany, Feb. 1966.

[12] Fischer, G.L., Jr., D.K. Pollock, B. Radack and M.E. Stevens, Eds., Optical Character Recognition (Spartan Books, Washington, D.C., 1962).

[13] Wirdzek, F.J., National Bureau of Standards, private communication, Oct. 28, 1966.

[14] Phonetics Laboratory, University College, London, Progress Report, July 1963, 36 p.

[15] Phonetics Laboratory, University College, London, Progress Report, Sept. 1965, 74 p.

[16] Fant, G., Acoustic Analysis and Synthesis of Speech with Applications to Swedish, reprint from Ericsson Technics No. 1, Sweden, 1959, 108 p.

[17] Risberg, A., Fundamental Frequency Tracking, Offprint, Proc. International Congress of Phonetic Sciences, 1961, pp. 228-231 (Mouton & Co.'s, Gravenhage, 1962).

[18] Pickett, J.M., Transmitting Speech Sounds by a Tactual Vocoder and by Lip-Reading, Rept. No. 27, 35 p. (The Speech Transmission Laboratory, Div. of Telegraphy-Telephony, The Royal Institute of Technology, Stockholm, Sweden, March 1963).

[19] Fant, G., Acoustic Theory of Speech Production, 323 p. (Mouton & Co.'s, Gravenhage, 1960).

[20] Fant, G. and K.N. Stevens, Systems for Speech Compression, Fortschritte der Hochfrequenz- technik 5, 229-262 (1960).

[21] Mártony, J., C. Cederlund, J. Liljencrantz and B. Lindblom, On the Analysis and Synthesis of Vowels and Fricatives, offprint from Proc. Fourth International Congress of Phonetic Sciences, 1961, pp. 208-213 (Mouton & Co., 1962).

[22] Liljencrantz, J., A Filter Bank Speech Spectrum Analyzer, Paper A-27, Fifth International Congress on Acoustics, Liège, Sept. 1965.

[23] Fant, G., K. Fintoff, J. Liljencrantz, B. Lindblom and J. Mártony, Formant Amplitude Measurements, J. Acoust. Soc. Amer. 35, 1753-1761 (1965).

[24] Tillman, H.G., G. Heike, H. Schelle and G. Ungeheuer, DAWID I- Ein Beitrag zur Automatischen 'Spracherkennung', Paper A-12, Fifth International Congress on Acoustics, Liège, Sept. 7-14, 1965, 4 p.

[25] Ungeheuer, G., Ein Einfaches Verfahren zur Akustischen Klassifikation von Sprechern, Paper A-17, Fifth International Congress on Acoustics, Liège, Sept. 7-14, 1965, 3 p.

[26] Tillman, H.G., Zur Klassifikation der Individuumgebundenen Merkmale am Sprachschall, Paper A-47, Fifth International Congress on Acoustics, Liège, Sept. 7-14, 1965, 4 p.

[27] Ungeheuer, G., R. Rupprath and L. Friedrich, Zur Entwicklung eines Verbund Systems von Periodizitätsanalysator (Tonhöhenschrieber) und Intersimeter, Paper J-11, Fifth International Congress on Acoustics, Liège, Sept. 7-14, 1965, 3 p.

[28] Mainka, W., Zur Ketektion Nichtlinear Verzerrter Signale, Fifth International Congress on Acoustics, Liège, Sept. 7-14, 1965, 4 p.

[29] Stock, D., Zum Problem der Periodizitätsdetektion im Sprachschall, Forschungsbericht 66/3, Institut für Phonetik und Kommunikatiunsforschung der Universität Bonn, 1966, pp. 1-26.

[30] Vieregge, W.H., Die Akustische Struktur der Plosivlaute, Forschungsbericht 66/3, Institut für Phonetik und Kommunikationsforschung der Universität Bonn, 1966, pp. 1-48.

[31] Kotten, K., Erzeugung und Analyse von Korrelationsfunktionen Deutscher Lautsignale, Forsehungsbericht 66/3, Institut für Phonetik und Kommunikationsforschung der Universität Bonn, 1966, pp. 1-16, appendices.

[32] Tillman, H.G., Über Kommunikative und Extrakommunikative Sprechergebundene Merkmale, Forschungsbericht 66/4, Institut für Phonetik und Kommunikationsforschung der Universität Bonn, Mar. 1966, 177 p.

[33] Kusch, H., Automatic Recognition of Spoken Numbers (Digits), NTZ Communications J. 4, 201-206 (1965). English Edition reprint, Telefunken, 8 p.

[34] Von Keller, T., Die Kennzeichnung von Sprachlauten durch Spektrum, Autokorrelationsfunktion und Nulldurghgangsabstände, Zur Erlangung des Akademischen Grades Cines Doktor-Ingenieurs von der Fakultät für Elektrotechnik der Technischen Hochschule Karlsruhe genehmigte Dissertation, Karlsruhe, June 1966, 85 p.

[35] Gamba, A., G. Palmieri and R. Sanna, Self-Learning in PAPA, Suppl. Nuovo Cimento 20, 146-147 (1961).

[36] Gamba, A., A Multilevel PAPA, Suppl. Nuovo Cimento 26, 176-177 (1962).

[37] Gamba, A., Remarks on the Theory of PAPA, Suppl. Nuovo Cimento 26, 169-175 (1962).

[38] Palmieri, G. and R. Sanna, Automatic Probabilistic Programmer/Analizer for Pattern Recognition, Estratto Rivista Methodos XII, 126 (1960).

[39] Gamba, A., The Papistor. An Optical PAPA Device, Suppl. Nuovo Cimento 26, 371-373 (1962).

[40] Gamba, A., G. Palmieri and R. Sanna, Preliminary Experimental Results with PAPA No. 2, Suppl. Nuovo Cimento 23, 280-284 (1962).

[41] Palmieri, G. and R. Sanna, A New PAPA Machine, Suppl. Nuovo Cimento, Series X, 23, 266-275 (1962).

[42] Palmieri, G., PAPA No. 3, Suppl. Nuovo Cimento, Series X, 30, 958-965 (1963).

[43] Borsellino, A. and A. Gamba, An Outline of a Mathematical Theory of PAPA, Suppl. Nuovo Cimento 20, 221-231 (1961).

[44] Palmieri, G., A Proposal for a High-Speed PAPA, Suppl. Nuovo Cimento 23, 276-279 (1962).

[45] Bertero, M., Errors and Self-Learning in PAPA, Suppl. Nuovo Cimento 23, 184-190 (1962).

[46] Ceccato, S., A Model of the Mind, Consiglio Nazionale delle Richerche, Rome, Italy, 1965, 61 p.

[47] Beltrame, R., Il Visore di Una Macchina Che Osserva e Descrive, Rassegna Electronica 1, No. 1, 10-16 (Apr. 1965).

[48] Ceccato, S., The Mechanization of Thought and Language Processes, preprint, Wiener Memorial Meeting, Genoa, Italy, Oct. 26-30, 1965, 15 p.

[49] Ceccato, S., Concepts for a New Systematics, Information Storage & Retrieval 3, 193-214 (Dec. 1967). [Delivered at the International Symposium on Relational Factors in Classification, Center of Adult Education, University of Maryland, June 8-11, 1966.]

[50] Ceccato, S., Operational Linguistics, Foundations of Language 1, 171-188 (1966).

[51] EURATOM - Thesaurus - Indexing Terms used within EURATOM's Nuclear Documentation System, Rept. No. EUR 500.e (Second Edition, Part One), Directorate Dissemination of Information, Center for Information and Documentation, European Atomic Energy Community, Brussels, Belgium, Dec. 1966, 90 p.

[52] Rolling, L.N., A Computer-Aided Information Service for Nuclear Science and Technology, EURATOM, CID, Brussels, Belgium, 1966, 29 p.

[53] Cremer, M., Integrated Training Policy for Documentalists on a National Level: New Trends in the Federal Republic of Germany, in Proc. FID Congress 1965, Vol. II, pp. 61-62 (Spartan Books, Washington, D.C., 1966).

[54] Schneider, K., Maschinelle Dokumentation in der Bundesrepublik Deutschland, Nachrichten für Dokumentation 16, 22-24 (1965).

[55] Bernhardt, R., Computer-Einsatz der Herstellung der Deutschen Biliographie, Nachrichten für Dokumentation 17, 23-30 (1966).

[56] Gundlach, R., HISDOC/HDS - Ein Dokumentationssystem zur Inhaltlichen Erfassung und Maschinellen Erschlifzong Historischer Sekundärliteratur, Munich, 1965.

[57] Brodda, B. and H. Karlgren, Citation Index and Similar Devices in Mechanized Documentation, Rapport Nr. 2 till Kungl. Statskontoret, SKRIPTOR, Stockholm, Sweden, May 27, 1965, 13 p.

[58] Allén, S., Report on Work in Computational Linguistics, paper presented at the Colloque Sur la Mécanisation et l'Automation des Recherdes Linguistiques, Prague, Czechoslovakia, June 7-10, 1966.

[59] Krallman, D., Statistische Methoden in der Stilistischen Textanalyse: Ein Beitrag zur Informationserschliefzung mithilfe Elektronischer Rechenmaschinen, Doctoral Dissertation, Philosophischen Fakultät der Rhenischen Friedrich-Wilhelms-Universität zu Bonn, Bonn, 1966, 249 p.

[60] Krallman, D., Allgemeine Benutzerorientierte Programmbibliothek für Linguistische Aufgabenstellungen, Forschungsbericht 66/2, Institut für Phonetik und Kommunikationsforschung, Universität Bonn, Bonn, 1966, 43 p.

[61] Krallman, D., T. Krumnack and H. Schelle, Kodierungssystem zur Verkartung und Maschindlen Verarbeitung beliebiger Texte, Forschungsbericht 66/2, Institut für Phonetik und Kommunikationsforschung, Universität Bonn, Bonn, 1966, 30 p.

[62] Schelle, H., CC-Automata and CF-Grammars, paper presented at the Colloquium on Algebraic Linguistics and Automata Theory, Jerusalem, Aug. 24-25, 1964, 21 p.

[63] Glasersfeld, E.v., A Project for Automatic Sentence Analysis, Rept. No. ILRS-T3, 640331, Istituto Documentazione dell'Associazion: Meccanica Italiana, Milan, May 1964, 11 p.

[64] Glasersfeld, E.v., Automatic English Sentence Analysis, Rept. No. ILRS-T14, 660930, Istituto Documentazione dell'Associazione Meccanic Italiana, Milan, Italy, Sept. 1966, 102 p.

[65] Glasersfeld, E.v., The Functions of the Articles in English, Centro di Cibernetica e di Attività Linguistiche, University of Milan, Feb. 1963, 21 p.

[66] Burns, J., English Prepositions in Machine Translation, Rept. No. ILRS-T9, 650115, Istituto Documentazione dell' Associazione Meccanica Italiana, Milan, Italy, Jan. 1965, 15 p.

[67] Glasersfeld, E.v., An Approach to the Semantics of Prepositions, Rept. No. ILRS-T12, 651115, Istituto di Documentazione dell' Associazione Meccanica Italiana, Milan, Italy, Dec. 1965, 16 p.

[68] Glasersfeld, E.v., 'MULTI STORE' - A Procedure for Correlational Analysis, Rept. No. ILRS-T10, 650120, Istituto Documentazione dell' Associazione Meccanica Italiana, Milan, Italy, Jan. 1965, 80 p.

[69] Glasersfeld, E.v., and B. Dutton, Supralinguistic Classification and Correlators, Rept. No. ILRS-T13, 660320, Istituto Documentazione dell' Associazione Meccanica Italiana, Milan, Italy, Mar. 1966, 30 p.

[70] Stickel, G., Automatische Textzerlegung und Registerherstellung, Rept. No. PI-11, Deutsches Rechenzentrum, Darmstadt, Federal Republic of Germany, Dec. 1964, 16 p.

[71] Meetham, A.R., Preliminary Studies for Machine Generated Index Vocabularies, Language and Speech 6, Pt. 1, (Jan.-Mar. 1963).

[72] Vaswani, P.K.T., Mechanized Storage and Retrieval of Information, Rev. Int. Doc. 32, 19-22 (1965).

[73] Wagner, S.W., Automatische Stichwortanalyse nach dem Rangkriterienverfahren, Doctoral Dissertation, Fakultät für Elektrotechnik der Technischen Hochschule Karlsruhe, Karlsruhe, Federal Republic of Germany, 1966, 116 p.

[74] United Nations Educational, Scientific and Cultural Organization, Symposium on Mechanized Abstracting and Indexing - Final Report; UNESCO/NS/209, Paris, Apr. 28, 1967, 6 p.

[75] The German Computing Centre - General Information, Deutsches Rechenzentrum, Darmstadt, Federal Republic of Germany, Feb. 1965, 13 p.

[76] Pilch, H., F. Schulte-Tigges, H. Seiler, G. Ungeheuer, Die Struktor formalisierter Sprachen, Deutsches Rechenzentrum, Darmstadt, Federal Republic of Germany, Oct. 1965, 38 p.

[77] Reul, H. and G. Miedel, Problems and Tools of Nonnumerical Data Processing at the German Computing Center, paper presented at IX European SHARE Conference, Grénoble, October 1965, Deutsches Rechenzentrum, Darmstadt, Federal Republic of Germany, 15 p., appendices.

[78] Görke, W., Selbstkorrigierende Decodierschaltungen für Hammingcodes mit Mindestabständen Zwei, Drei und Vier, NTZ 6, 352-367 (1966).

[79] Caracciolo di Forino, A., Linguistic Problems in Programming Theory, Proc. IFIP 1965, pp. 223-228.

[80] Caracciolo di Forino, A., String Processing Languages and Generalized Markov Algorithms, Centre Studi Calcolatrici Elettroniche Universitá di Pisa, n.d., 26 p.

[81] Carracciolo di Forino, A., Some Preliminary Remarks on Theoretical Pragmatics, Commun. ACM 9, 226-227 (Mar. 1965).

[82] Caracciolo di Forino, A., Special Programming Languages, Centro Studi Calcolatrici Elettronishe, Universitá di Pisa, Pisa, Italy, 1965, 21 p.

[83] Caracciolo di Fornio, A., An Elementary Constructive Theory of Discrete Probabilistic Processes and Parallel Concurrent Processes, Centro Studi Calcolatrici Elettronische, Universitá di Pisa, Pisa, Italy, Oct. 1965, 32 p.

[84] Caracciolo di Forino, A. and G. Molnar, Sistemi Dinamici e Linguaggi di Simulazione, Centro Studi Calcolatricci Elettronische, Universitá di Pisa, Pisa, Italy, May 1966, 23 p.

[85] Caracciolo di Forino, A., L. Spanedda and N. Wolkenstein, PANON - 1B: A Programming Language for Symbol Manipulation, paper presented at the SIC-SAM Symposium, Washington, D.C., Mar. 29-31, 1966, Centro Studi Calcolatrici Elettroniche, Universitá di Pisa, Pisa, Italy, 1966, 21 p.

[86] Caianiello, E.R. and C. Crocchiolo, Programma 'Procuste' per l'Analisi di Linguaggi Naturali, Calcolo 2, fasc. 1, 83-101 (Jan.-Mar. 1965).

[87] Caianiello, E.R., On the Analysis of Natural Languages ('Procrustes' Program), Proc. 3rd All Union SSR Congress on Cybernetics, Odessa, Sept. 1965, Rept. No. SPT/Doc/E.R.C. 1, 12 p.

[88] Caianiello, E.R., Outline of a Theory of Thought-Processes and Thinking Machines, Consiglio Nazionale delle Richerche, Rome, 1965, 27 p.

NBS TECHNICAL PUBLICATIONS

PERIODICALS

JOURNAL OF RESEARCH reports National Bureau of Standards research and development in physics, mathematics, chemistry, and engineering. Comprehensive scientific papers give complete details of the work, including laboratory data, experimental procedures, and theoretical and mathematical analyses. Illustrated with photographs, drawings, and charts.

Published in three sections, available separately:

- **Physics and Chemistry**

Papers of interest primarily to scientists working in these fields. This section covers a broad range of physical and chemical research, with major emphasis on standards of physical measurement, fundamental constants, and properties of matter. Issued six times a year. Annual subscription: Domestic, $6.00; foreign, $7.25*.

- **Mathematical Sciences**

Studies and compilations designed mainly for the mathematician and theoretical physicist. Topics in mathematical statistics, theory of experiment design, numerical analysis, theoretical physics and chemistry, logical design and programming of computers and computer systems. Short numerical tables. Issued quarterly. Annual subscription: Domestic, $2.25; foreign, $2.75*.

- **Engineering and Instrumentation**

Reporting results of interest chiefly to the engineer and the applied scientist. This section includes many of the new developments in instrumentation resulting from the Bureau's work in physical measurement, data processing, and development of test methods. It will also cover some of the work in acoustics, applied mechanics, building research, and cryogenic engineering. Issued quarterly. Annual subscription: Domestic, $2.75; foreign, $3.50*.

TECHNICAL NEWS BULLETIN

The best single source of information concerning the Bureau's research, developmental, cooperative and publication activities, this monthly publication is designed for the industry-oriented individual whose daily work involves intimate contact with science and technology—*for engineers, chemists, physicists, research managers, product-development managers, and company executives.* Annual subscription: Domestic, $3.00; foreign, $4.00*.

*Difference in price is due to extra cost of foreign mailing.

NONPERIODICALS

Applied Mathematics Series. Mathematical tables, manuals, and studies.

Building Science Series. Research results, test methods, and performance criteria of building materials, components, systems, and structures.

Handbooks. Recommended codes of engineering and industrial practice (including safety codes) developed in cooperation with interested industries, professional organizations, and regulatory bodies.

Special Publications. Proceedings of NBS conferences, bibliographies, annual reports, wall charts, pamphlets, etc.

Monographs. Major contributions to the technical literature on various subjects related to the Bureau's scientific and technical activities.

National Standard Reference Data Series. NSRDS provides quantitative data on the physical and chemical properties of materials, compiled from the world's literature and critically evaluated.

Product Standards. Provide requirements for sizes, types, quality and methods for testing various industrial products. These standards are developed cooperatively with interested Government and industry groups and provide the basis for common understanding of product characteristics for both buyers and sellers. Their use is voluntary.

Technical Notes. This series consists of communications and reports (covering both other agency and NBS-sponsored work) of limited or transitory interest.

CLEARINGHOUSE

The Clearinghouse for Federal Scientific and Technical Information, operated by NBS, supplies unclassified information related to Government-generated science and technology in defense, space, atomic energy, and other national programs. For further information on Clearinghouse services, write:

Clearinghouse
U.S. Department of Commerce
Springfield, Virginia 22151

Order NBS publications from:
Superintendent of Documents
Government Printing Office
Washington, D.C. 20402

U.S. DEPARTMENT OF COMMERCE
WASHINGTON, D.C. 20230

OFFICIAL BUSINESS

POSTAGE AND FEES PAID
U.S. DEPARTMENT OF COMMERCE