IBM Language Processing Systems

IBM Data Systems Division, Kingston, New York

December 15,1964

CONTENTS

Heading Page

Introduction

IBM is engaged in an intensive research and development program in the area of non-numeric information processing, with current emphasis on automatic language translation and information retrieval. The program originated in the Systems Applications and Technology Research Department at the Thomas J. Watson Research Center, and is being implemented by the Data Systems, General Products, and Data Processing Divisions. This combined effort involves work on all aspects of optical character recognition and lexical data processing, including multi-font character-recognition research, basic and applied linguistics, lexicography, programming, information-retrieval studies, and experimentation. The program also includes machine organization, system design, logical design, advanced applications studies, circuit design, hardware development, system construction, experimentation, evaluation, and improvement.

During the past six years, IBM developed and has successfully operated the AN-GSQ-16 Mark II language translator for the Air Force. IBM also developed the Research Language Processor-3 (RLP-3) (figure 1) for machine translation of languages at the 1964-1965 World's Fair, and is currently working on an advanced model. The RLP-3 design embodies the very latest concepts, both in terms of the hardware design and software organization, that have evolved in the field of language processing.

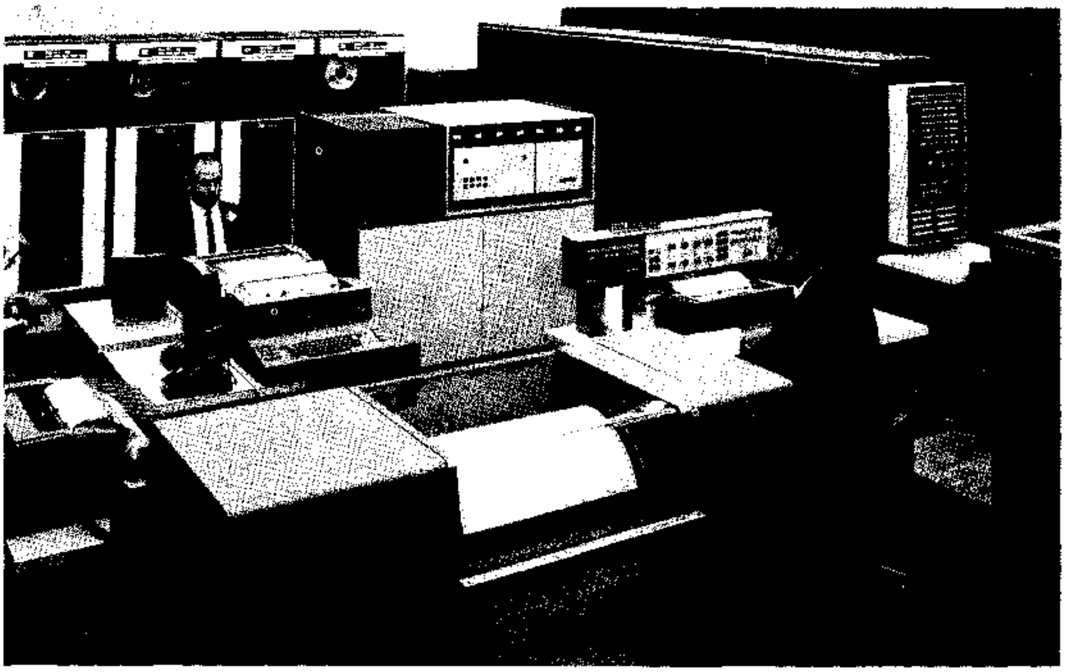

Figure 1. Research Language Processor

Although all original developments were concerned with Russian-to-English language translation, the current system can translate languages in two ways:

1. It can perform machine translation by selecting the program and dictionary of the desired language. Russian-English dictionary programs are now available and will operate at through-put speeds of approximately 25 words per second. Chinese-English, French-English, English-French, and others are being studied.

2. It can perform machine transcription of stenographic tape which contains the translated information. Steno/English-English dictionary programs are now available and will operate at through-put speeds of approximately 60 words per second.

Machine Translation

The foreign language text is recorded on an electric typewriter with a 6-channel paper tape. The translation process is completely automatic, being accomplished in a single left-to-right pass through the input, one sentence at a time. The translation of each word is retrieved from the automatic dictionary and grammatically analyzed for possible modification. The range of the analysis encompasses about three words on either side of the word being translated. When the end of a record is reached, the translation is shifted out of the machine and printed out on the high-speed printer or stored on magnetic tape. Post-editing, though not required for many applications, may be performed to obtain more polished translations, depending on the user's requirements.

Russian - English Translation

The Russian-English automatic dictionary now in use contains approximately 170,000 entries, each of which has a Russian field, an English field, and a grammatical tag. Approximately 3 million Russian words can be handled with these stems because of the derivational capabilities of the operational program. Since the dictionary is organized according to the longest-match principle, phrases may be included in the dictionary and will be matched as single units rather than on a word-by-word basis. Nominal, verbal, adjectival, and adverbial phrases are also given grammatical tags, and, when appropriate, they are treated in the analysis as nouns , verbs, adjectives, or adverbs.

With a system utilizing a 1401 as an I/O controller, the through-put rate of the Russian-English language translator is approximately 25 words per second.

Steno Transcription

A stenotypist may be employed to record translations when information is transferred orally. Stenotype machines are used to record translated information in the user's language. The resultant record comprises the basic input to the Steno Mark Reader, which is capable of .directly reading the printed stenographic marks. The Steno Mark Reader encodes each printed character and stores the information on a magnetic tape unit. The reading rate of the mark sensor is sufficiently high to enable it to handle the simultaneous output of a large number of steno keyboards.

Steno/English - English Transcription

In the current system, automatic transcription of the steno input data into English is accomplished by successive table lookup operations, utilizing an automatic dictionary program. Most of the dictionary entries are stored in the Photo Store Unit. However, selected entries that are retrieved with relatively high frequency are stored in the core storage of the Language Processor. The program, in turn, examines each input item by comparing it to entries in the dictionary. When a match is obtained, the English transcription is assembled and various address registers are modified to specify the next table lookup operation. Whenever a potentially ambiguous steno item is encountered, the program is directed to resolve the ambiguity by examining the local context surrounding the item and comparing it with predetermined algorithms. If the ambiguity cannot be resolved, the alternate transcriptions are read out and flagged by separating them with a diagonal mark. Automatic editing currently consists of (1) automatic deletion upon recognition of a deletion code used by the input operator; (2) automatic insertion of missing punctuation under certain conditions, such as space, capitalization, and periods before new paragraphs; (3) automatic rectification of common speech anomalies, such as "the the," etc. Error correction is accomplished by allowing several lookups for the same item before transliteration; this compensates for many nonpermanent machine errors and the use of special dictionary entries for common, possible misspellings of frequently occurring words.

With a system configuration utilizing a 1401 as an I/O controller, a throughput rate of approximately 60 words per second can be maintained. At this rate, the automatic steno system can accommodate more than 24 stenotypists working full time at an average rate of about 150 words per minute.

System Configuration

The Language Processing System contains the following units (see figures 2 and 3):

a. Standard 1401 Data Processor for data communication and I/O operations,

b. A 7256 Photo Store Memory for dictionary operation based on rapid access to its large storage.

c. A special-purpose 7266 Lexical Processor for high-speed translation.

d. A 1903 Paper Tape Reader for input information.

e. A 1403 high-speed Printer for output information.

f. A 729 Magnetic Tape Drive for I/O information storage, program modification , etc.

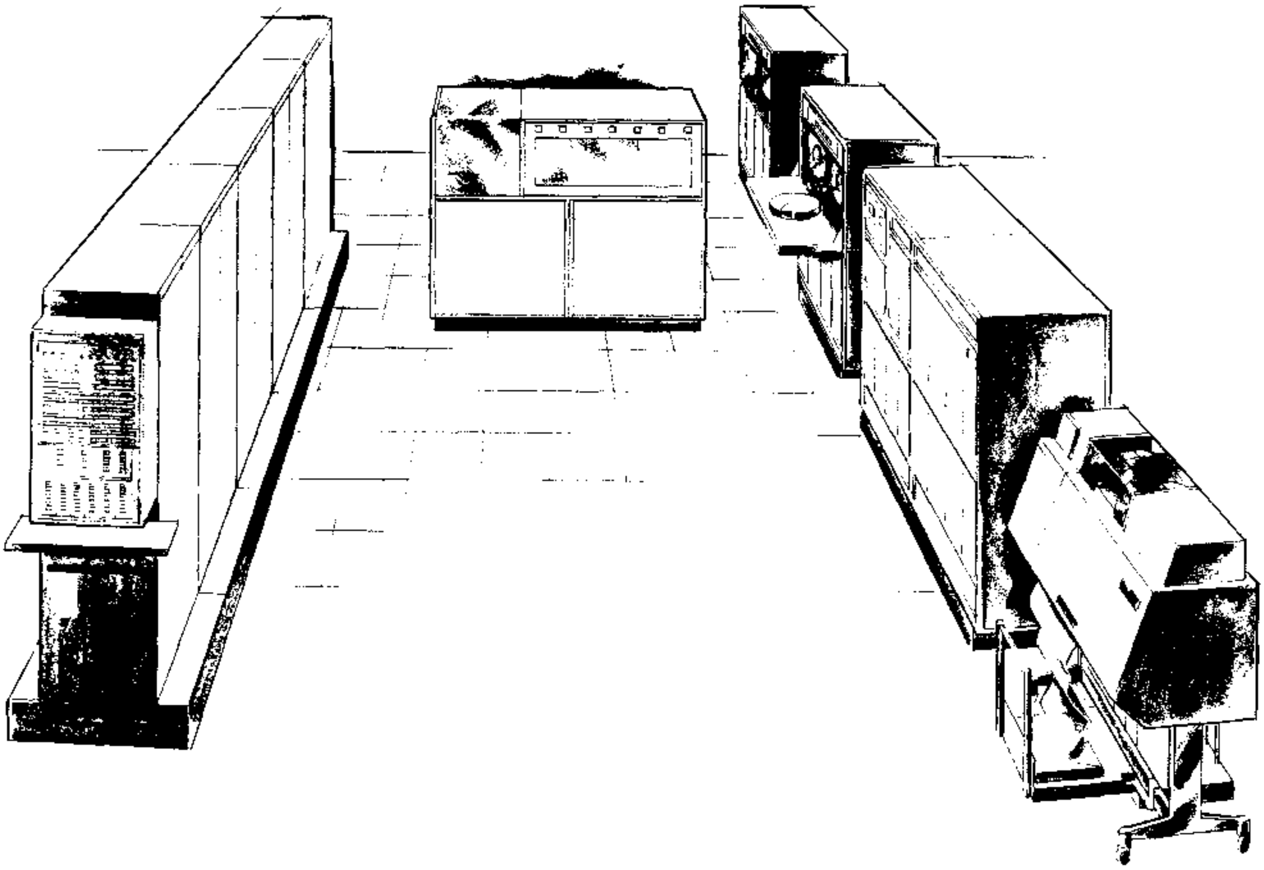

Figure 2. Typical Translation Equipment

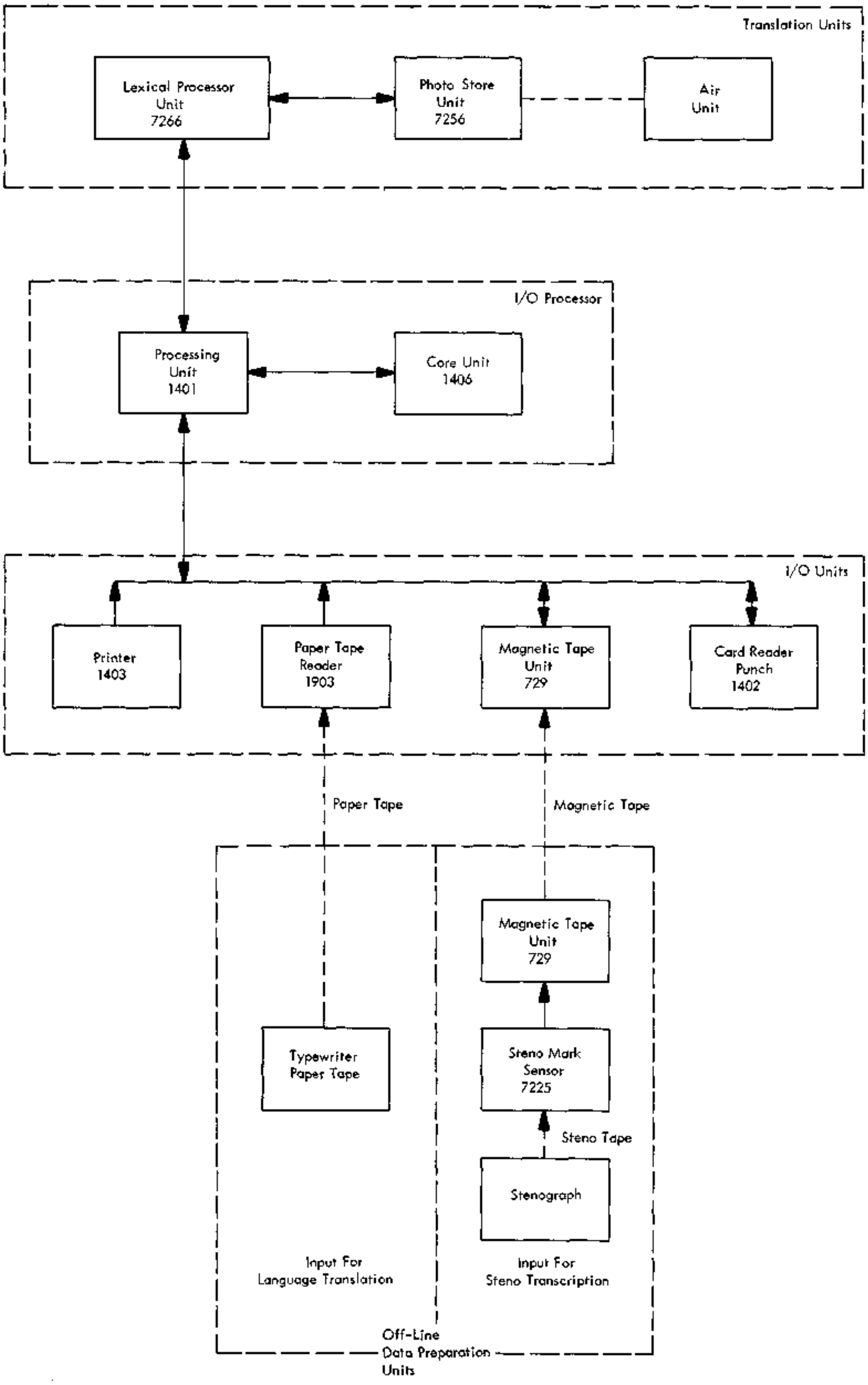

Figure 3. Translation/Transcription System

Equipment Description

The Language Processing System contains standard IBM commercial products, such as the IBM 1401, and specially designed equipment. Since descriptive material is available for the IBM 1401 and any other general-purpose data processing system, only the special equipment and associated I/O units are described herein. The special equipment includes a Lexical Processor and a Photo Store Unit.

Lexical Processor Unit

The Lexical Processor Unit (LPU) is a highly functional logical device incorporating a flexible program with fixed lexical controls. The LPU is an extremely efficient device for manipulating characters and performing table lookup operations with variable word lengths. The lexical logic is not limited to a particular language and/or dictionary. The LPU consists of a console, core unit, and lexical logic. The integrated 160,000-character, 2.5-/usec core storage contains high-frequency word tables, errata, addenda, control entries, and photo store indexing.

The LPU controls the translation process by executing instructions, compiling translated information, and correlating data and directing it to the Photo Store Unit, core storage, or the general-purpose data processor.

Photo Store Unit

The Photo Store Unit (PSU) supplements the high-speed working core storage of the Lexical Processor Unit by providing high capacity dictionary storage with moderate access time and high-speed reading.

The tables and dictionary entries contained in the Photo Store are compiled by the combined efforts of programmers, linguists, and lexicographers. The dictionary entries contain the source language to be translated and associated target language translation; these entries also contain operating instructions for the Lexical Processor Unit.

A complete dictionary, nominally containing 170,000 entries, is ordered alphabetically, and each of the characters is assigned a unique binary code (including the instructions). The sequence of binary numbers is thereby automatically ordered from low to high numbers and the longest entry of each alphabetic group is assigned the highest number or weight. Matching is conducted in such a way • that the longest possible string of characters (including spaces) is selected in each lookup operation. This has been likened to looking at pages in a normal dictionary from the bottom of the page upwards. Thus, a match on JACKHAMMER would not be stopped prematurely by matching on JACK, and a match on BILL OF RIGHTS would not be stopped after having matched on BILL.

In the PSU, the storage medium is a mylar disk (figure 4) containing approximately 130 million bits of photographically recorded information. The information is on about 2000 concentric tracks concentrated within a 3/4-inch band near the outside edge of the disk. The bits are grouped into 8-bit characters: six data bits, one control bit, and one check bit.

The binary information is read by rotating the disk at approximately 2400 rpm and focusing a fine light beam from a cathode-ray-tube source through the disk by means of a movable lens. The lens is positioned by a movable arm operating in a closely controlled servo loop. The latter is capable of maintaining the light beam on a single track or of switching from track to track, as necessary, upon command from the Lexical Processor logic. The light beam is passed through the binary information (alternate light and dark marks) on the disk onto a photomultiplier tube.

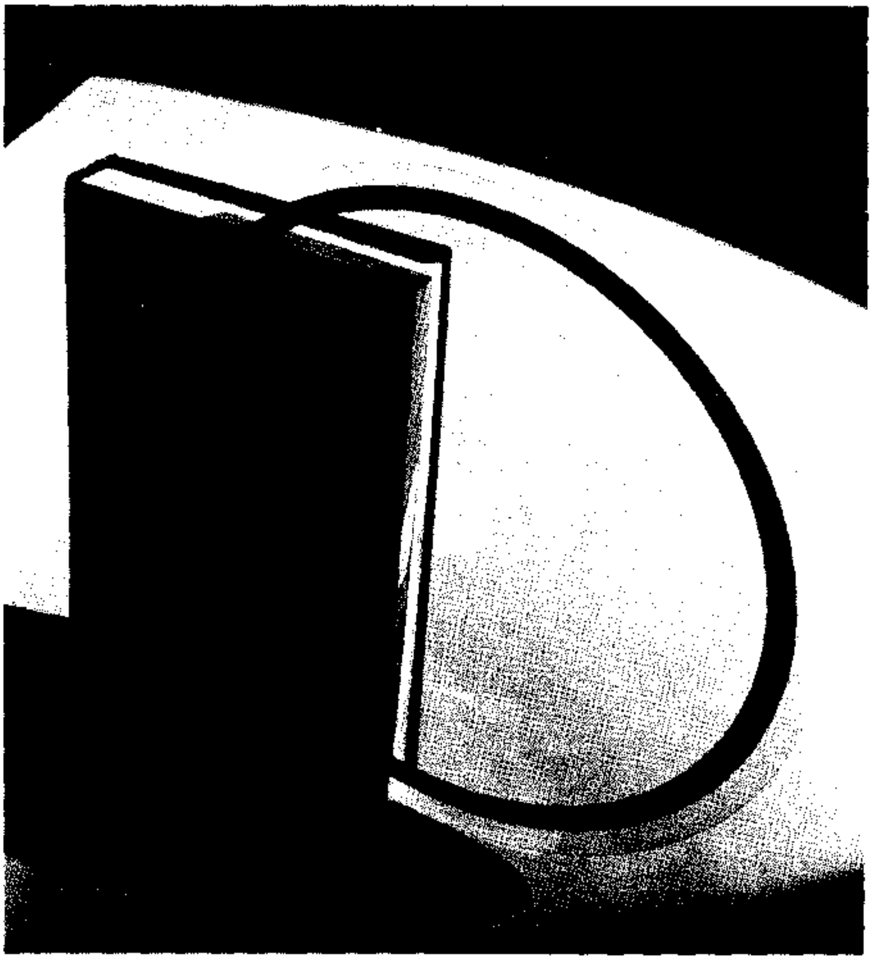

Figure 4. Photoscopic Disk

Input/Output Units

1903 Paper Tape Reader

The 1903 Paper Tape Reader is essentially a buffered version of the IBM 1011 Paper Tape Reader. It reads paper tape at a maximum rate of 500 characters per second. The paper tape can be chad or chadless; in the three common widths: 11/16 inch (5-track), 1 inch (8-track), or 7/8 inch (6- and 7-track); and in strips, reels, or rolls that feed from the center. Spoolers and controls are provided to automatically unwind and wind the tape reels.

The flexibility of the 1903 is increased by a control panel where characters read from the tape can be converted to a different character coding, can be omitted, or where several paper tape codes can be assigned to the same BCD coding. An odd-parity check code is automatic for paper tape codes using odd-bit parity.

1403 Printer

The 1403 Printer is well known for its ease of operation and its high-speed quality output in multiple copy. Horizontal spacing is 10 characters to the inch; vertical line spacing is six to eight lines to the inch, under operator control. The 1403 can print 48-character alphabets (approximately 600 1pm) or 80-character alphabets (approximately 389 1pm) to handle upper-case and lower-case characters and special characters for increased readability.

Typewriter Paper Tape Unit

The electric typewriter-paper tape unit consists of the following:

a. A 2-case typist keyboard (one case Cyrillic, one case Roman) (figure 5).

b. A 2-case print mechanism for hard copy (as above).

c. Paper tape punch unit.

d. Paper tape reader unit.

The paper tape punch is compatible with the 1903 Paper Tape Reader. The machine operates at approximately 12 to 15 characters per second.

The typewriter is a standard electric type with 88 characters, or 51 keys, and with special provisions for automatic carriage return during tape reading. See figure 6.

Steno Mark Reader

The Steno Mark Reader optically scans steno paper tape, detects characters appearing on the tape, converts the steno characters to binary- and decimal-coded characters, and records the information on magnetic tape.

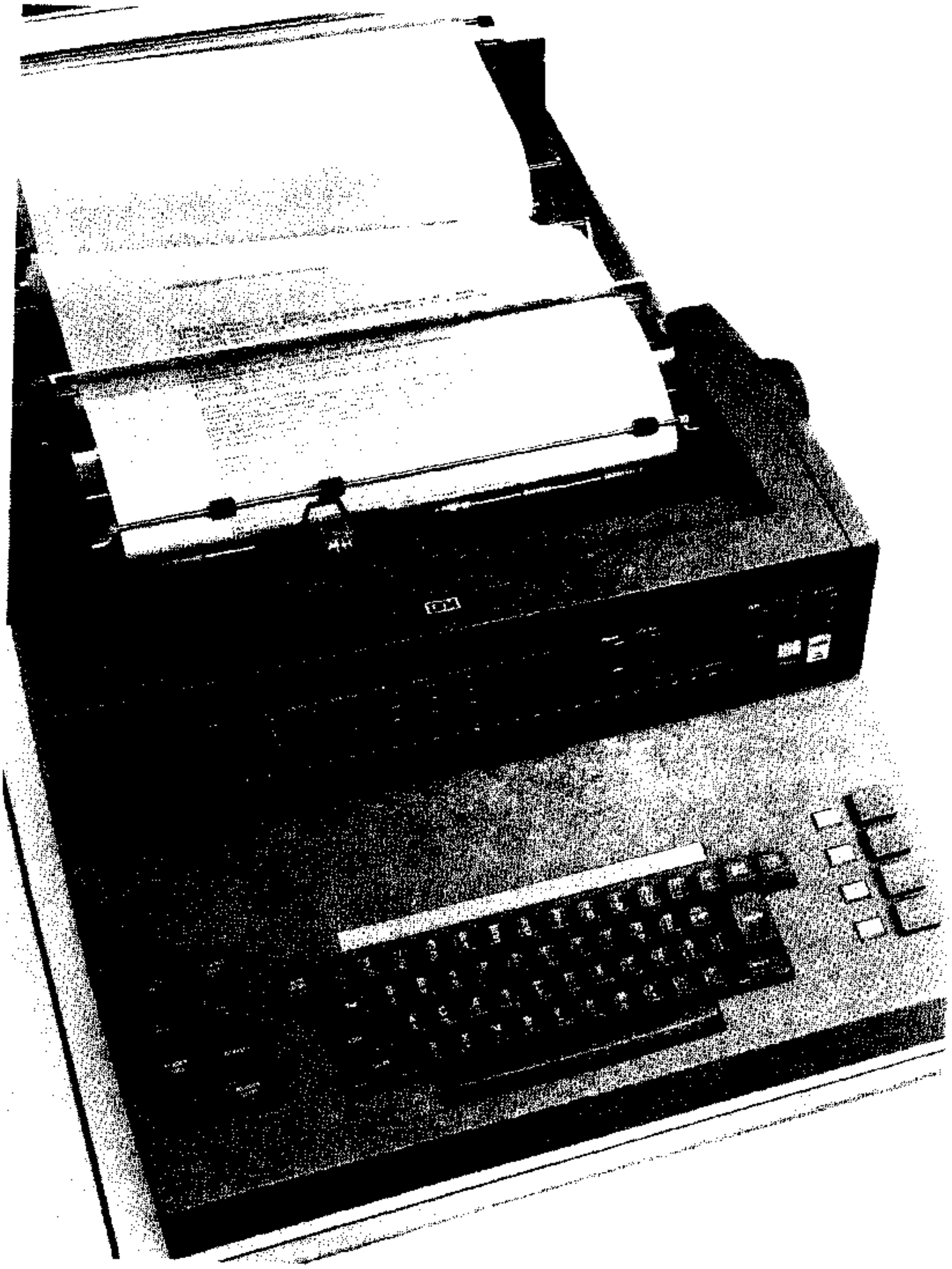

Figure 5.   Printer Keyboard

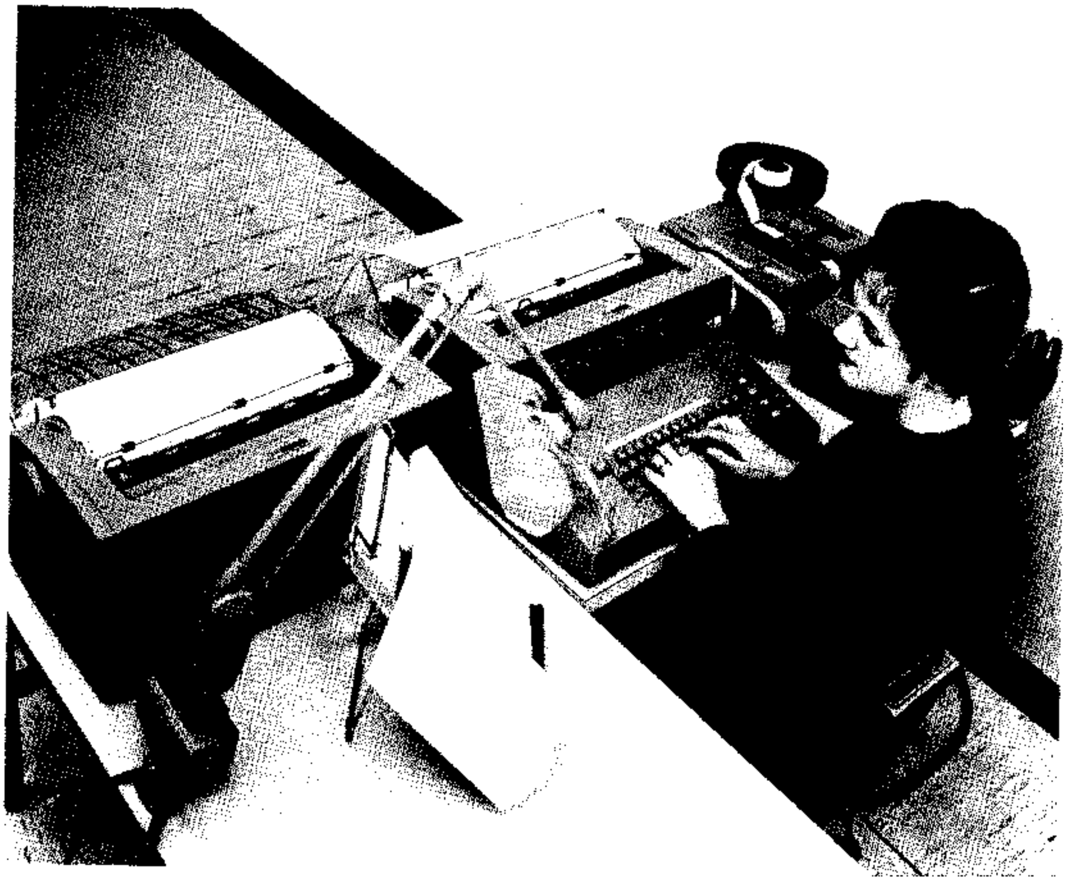

Figure 6. IBM 1050 Data Communications Terminal